Project acronym:	BYTE
Project title:	Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities
Grant number:	619551
Programme:	Seventh Framework Programme for ICT
Objective:	ICT-2013.4.2 Scalable data analytics
Contract type:	Co-ordination and Support Action
Start date of project:	01 March 2014
Duration:	36 months
Website:	www.byte-project.eu

Deliverable D1.3:
Big data initiatives

Author(s):	Rachel Finn, Anna Donovan and Kush Wadhwa, Trilateral Research and Consulting Lorenzo Bigagli, National Research Council of Italy José María García, Semantic Technology Institute (STI) – University of Innsbruck

Dissemination level:	Public
Deliverable type:	Final
Version:	1
Submission date:	31 October 2014

Table of Contents

EXECUTIVE SUMMARY

This deliverable provides an examination of ten big data initiatives that are useful examples of projects that involve big data, and highlight where they make societal, economic and technological developments, especially as a result of the collaboration between a number of different stakeholders. Understanding these initiatives assists us in gaining an insight into the current and evolving big data landscape, which is integral to addressing the issues raised by the BYTE project. For this report, we view "initiatives" to mean programmes seeking to implement courses of action. We examine initiatives undertaken by governments, commercial organisations and other organisations in a number of different sectors in order to identify what sorts of big data practices are currently in evidence, and what these different stakeholders seek to gain from these projects. As such this report will discuss the following 10 big data initiatives from within the pubic and private sectors:

- European Centre for Nuclear Research (CERN) – Worldwide LHC Computing Grid;
- US Big Data Research and Development Initiative;
- Australian Government Public Service Big Data Strategy;
- UK Data Service;
- eBay Inc. Big Data Analytics Programme;
- UN Global Pulse Initiative
- European Space Agency Big Data Initiative;
- European Bioinformatics Institute;
- Open data platforms/ vendors, including: Teradata; and
- DeCODE– Genetics, Iceland

In Chapter 2, we describe each of these initiatives and consider how their conceptualisation of big data, their policies and practices and their strategies for managing big data can inform the activities of BYTE. We examine what big data means in each of these initiatives and what stakeholders are involved. Finally, we look for technological, legal and ethical issues as well as economic and social developments and externalities and consider how these might provide preliminary data for BYTE.

The analysis results in three main conclusions for BYTE. First, the analysis challenges the traditional and long-accepted basic definition based on the 3Vs originally proposed by Gartner. All of these initiatives are dealing with large volumes of data. However, most are only dealing with either high velocity or data of a large variety. Very few are dealing with all three. Furthermore, this investigation also finds that the conceptualisation of high velocity data varies significantly based on the type of data being utilised. As such, BYTE needs to continue to consider the definition of "big data", and the extent to which it might be contextually specific, throughout the project. It suggests that the project may need to adopt a fluid definition of big data that is flexible and adaptable to different sectors.

Second, all the aforementioned big data policy initiatives involve cross-sector and cross-agency collaborations, business partnerships and similar emerging relationships. This demonstrates that form a policy perspective, the benefits of big data would best be achieved through different stakeholders working collaboratively. This is also one of the key objectives of BYTE and the Big data community that BYTE will form, and the policy push to create similar networks demonstrates that such networking and community building is understood to represent good practice in responsible innovation.

Finally, and most importantly, the aforementioned big data stories detail initiatives that have been implemented to utilise big data to produce positive results, impacts and externalities and provide initial information about how some big data initiatives are addressing potential negative externalities. Many of the initiatives examined here are examining innovative ways to protect privacy, prevent discrimination and ensure the protection of personal data. In addition, other initiatives demonstrate that significant economic gains can be realised through the leveraging of big data either to create profit, create jobs, develop new skills or support innovation. Finally, the analysis highlights how big data can be used for social gains, including evidence-based policy-making and the protection of citizen safety, health and well-being. Each of these aspects will be brought forward into the further work of BYTE as the project develops.

1 INTRODUCTION

1.1 OVERVIEW

The aim of this deliverable is to provide an examination of ten big data initiatives that are useful examples of projects that involve big data, and highlight where they make societal, economic and technological developments, especially as a result of the collaboration between a number of different stakeholders. Big data initiatives are evident in both the public and private sectors, and they are becoming increasingly prevalent due to advances in computing and data science that now make it possible to process and analyse big data in real-time and reap the insights from big data analytics. Consequently, there are an increasing number of big data initiatives being developed in Europe and around the world, some of which are discussed below.

Understanding these initiatives assists us in gaining an insight into the current, and evolving big data landscape, which is integral to addressing the issues raised by the BYTE project. These initiatives produce positive and negative externalities as a result of the technologies and practices that are implemented as important features of the initiative. We also note that, for the most part, the data utilised for these initiatives accords with the definitions of big data, as outlined at Deliverable 1.1 of BYTE, but namely with the references to the long accepted definition of big data as data comprising of the 3VS – Volume, Velocity and Variety.[1] The initiatives discussed below indicate that governments, companies and organisations in a variety of industries are committed to deriving real value from advanced analytics. These initiatives are made possible because public and private sector organisations and companies have vast amounts of data at their disposal that has been passively generated as by-products of people's everyday use of technologies and the information people willingly communicate about themselves on the web.[2] For this report, we view "initiatives" to mean programmes seeking to implement courses of action. As such this report will discuss the following 10 big data stories from within the pubic and private sectors:

- European Centre for Nuclear Research (CERN) – Worldwide LHC Computing Grid;
- US Big Data Research and Development Initiative;
- Australian Government Public Service Big Data Strategy;
- UK Data Service;
- eBay Inc. Big Data Analytics Programme;
- UN Global Pulse Initiative
- European Space Agency Big Data Initiative;
- European Bioinformatics Institute;
- Teradata; and
- DeCODE– Genetics, Iceland

These initiatives are taken from across a range of sectors and are indicative of the increasing implementation of initiatives specific to big data, its related technologies and practices. These stories also represent a development of thinking associated with big data and how big data and analytics can be applied across a number of departments and units within organisations and companies to drive revenue growth, increase productivity, lead to informed policy and

[1] Commonly cited as the Gartner definition: Laney, Doug, "3D Data Management: controlling Data Volume, Velocity and Variety", *META Group Application Delivery Strategies,* 6 February 2011.
[2] United Nations Global Pulse, "About", no date. http://www.unglobalpulse.org/about/faqs

decision-making and assist in humanitarian efforts and scientific discoveries, as well as leading to innovation in technologies and the growth of a big data specific workforce.

In terms of business and commercial development, with the help of data scientists, a new generation of business leaders and collaborators help marketers combine the art of creativity with the science of numbers to help drive insight and business results.[3] Big data initiatives are turning data into revenue but also producing benefits for society. Socio-economic benefits can be seen as flowing from the Australian and US government initiatives that promote data analytics as a key resource for the development of policies for governments and society. Value and consumer benefits are witnessed benefits resulting from eBay Inc.'s big data analytics programme, and the business models and big data analytics technologies promoted through Teradata. Science and society are capturing the benefits of the CERN initiative, the ESA big data initiative, and the European bioinformatics initiative. Therefore, big data initiatives are being implemented across a variety of sectors because of the emerging and recognised benefits of big data analytics. This is despite the challenges faced by big data actors, who are increasingly incorporating solutions to the issues and challenges faced into strategies and policies for initiatives. This is an indication that some strategies for addressing potential negative externalities such as legal and ethical issues raised by the big data technologies and practices are having moderate success in enabling these important and valuable initiatives.

In the pages that follow we examine each of these initiatives and consider how their conceptualisation of big data, their policies and practices and their strategies for managing big data can inform the activities of BYTE. We examine what big data means in each of these initiatives and what stakeholders are involved. Finally, we look for technological, legal and ethical issues as well as economic and social developments and externalities and consider how these might provide preliminary data for BYTE.

[3] Arthur, Lisa, "What Does it Take to Turn Big Data Into Big Dollars?", *Forbes*, 27 March 2012. http://www.forbes.com/sites/lisaarthur/2012/03/27/what-does-it-take-to-turn-big-data-into-big-dollars/

2 BIG DATA INITIATIVES

2.1 OVERVIEW

This report provides an overview of ten big data stories. They provide insight into how and why businesses and organisations are aligning themselves with big data in order to capture an array of recognised and emerging benefits that flow from big data analytics. These initiatives are important in terms of boosting the big data economy and supporting future developments within the big data economy. Whilst the initiatives reveal a number of positive externalities associated with their implementation, they do present challenges for businesses and governments alike, including social, ethical and legal issues. However, a number of the initiatives focus on minimising the impact of negative externalities as part of the policy or strategy in order to ensure their longevity. This provides BYTE with some useful inroads to begin identifying and considering potential positive and negative externalities raised by big data practices, as well as how these might be amplified or minimised, respectively. These initiatives also reveal the importance of involving multiple stakeholder groups as well as encourage an attention to how big data definitions manifest in practice.

2.1.1 CERN Worldwide LHC Computing Grid

The European Centre for Nuclear Research has been at the forefront of computing innovations since its foundation in 1954. In consequence, CERN has always been dealing with the challenges associated with the interaction between developments in data collection and analysis, and "big data" is a challenge that they have been grappling with since "big" meant gigabytes rather than petabytes or exabytes. The massive amounts of data generated by the Large Hydron Collider at CERN has required the organisation to develop "the most sophisticated data-taking and analysis system ever built"[4] through which to store and analyse that data. The development of this Worldwide LHC Computing Grid (WLCG) has encountered a number of externalities, including stakeholder integration, capitalising on innovation and technological innovations that make the project relevant for BYTE.

The Worldwide LHC Computing Grid (WLCG) is a distributed storage and analysis computing network. Essentially, the system was developed to successfully handle the huge volume of data generated by the Large Hydron Collider (LHC). The LHC has 150 million sensors[5] and generates approximately 30 petabytes of data each year[6]. Existing computing infrastructures are not able to store and analyse this data, and so in 2001 specialists at CERN decided that the solution was to "divide and conquer".[7] The resulting WLCG is made up of 170 computing centres in 40 countries distributed among six continents.[8] It enables "the seamless sharing of resources around the globe"[9] that hides "the complexity and location of

[4] Brumfiel, Geoff, "Down the Petabyte Highway", Nature, Vol. 469, 20 January 2011, p. 282-283 [p.282]. http://www.mcs.anl.gov/uploads/cels/files/news/2011/469282a.pdf
[5] Shiers, Jamie, "Longterm data preservation in HEP", *Best Practices for Data Management & Sharing*, Joint Research Centre Workshop, 14-15 April 2014. https://indico.cern.ch/event/313634/.
[6] Worldwide LHC Computing Grid, "Welcome to the Worldwide LHC Computing Grid", *CERN*, 23 October 2014. http://wlcg.web.cern.ch
[7] Brumfiel, Geoff, "Down the Petabyte Highway", *Nature*, Vol. 469, 20 January 2011, p. 282-283 [p.282]. http://www.mcs.anl.gov/uploads/cels/files/news/2011/469282a.pdf
[8] Worldwide LHC Computing Grid, op. cit., 2014.
[9] Smith, Tim, "Big Data", *TED-Ed*, 2014. http://ed.ted.com/lessons/exploration-on-the-big-data-frontier-tim-smith

both CPU and storage resources behind consistent interfaces".[10] The grid is divided into three "tiers". Tier 0 is located at CERN, and it is where data is collected, recorded and where the initial "cleaning" is accomplished. The data is then distributed among the 11 Tier 1 centres. This is where data is stored permanently, where it is processed a second time and where some initial analyses take place. Close analyses and data simulations are run using the approximately 2000 Tier 2 centres.[11] These three tiers that make up the computing grid will be responsible for managing the 100 petabyte archive that CERN currently has, as well as the potential .5 exabytes that CERN is expected to be producing annually by 2027.[12]

In addition to being high volume, the CERN data is also high velocity in two ways. First, the system requires a high velocity transfer of information from Tier 0 to the Tier 1 and Tier 2 centres. These data transfers, particularly for Tier 0 to Tier 1, must occur at a rate of between 50 – 200 MB per second 24 hours a day for approximately 100 days per year.[13] Second, the analysis of such large data sets must also occur at high velocity. Specifically, the WLCG is intended to be a "reliable petascale computing services that allow[s] data to be turned into discoveries in record time."[14] While the "Grid enables scientists to run vast analyses that would push the world's most powerful supercomputers to the edge", the volume of the data is evidenced by the fact that this sophisticated and innovative system still requires days to run analyses for individual scientists.[15] This is what "high velocity" means in an era of big data analysis.

What makes the computing grid interesting is that it is purely a support service for the LHC experiments, and there is no independent R&D component.[16] However, the system itself is an essential part of the LHC experiments and findings, including the Higgs discovery, as it would not have been possible to analyse the LHC data without the Grid.[17] Thus, the WLCG makes the analysis of big data, and the associated externalities possible. In relation to externalities, the project has had a significant economic and innovation impact. Specifically, Shiers estimates that the $4Bn investment in LHC technology has resulted in approximately $50Bn in returns through education, spin off technologies and advances in computing.[18] Second, the project had to meaningfully and continuously manage the computing centres themselves to get all of the "institutes and individuals, all with existing, sometimes conflicting commitments, to work together".[19] As such, stakeholder values and motivations had to be considered. The project also had to tackle challenges related to hardware acquisition and training. Finally, the project was developed over more than a decade. The WLCG was initially designed in 2001 and was not ready for use until 2012.[20] This demonstrates that big data projects must take a long view to ensure that they remain relevant during their life cycle and to ensure that they account for advances in technology during their development.

---

[10] Shiers, Jamie, "The Worldwide LHC Computing Grid (Worldwide LCG)", *Computer Physics Communications*, Volume 177, Issues 1–2, July 2007, Pages 219-223, p. 220. http://jamie.home.cern.ch/jamie/ccp2006.pdf

[11] Shiers, Jamie, "Hunting the Higgs: Using the Worldwide LHC Computing Grid", *Best Practices for Data Management & Sharing*, Joint Research Centre Workshop, 14-15 April 2014. https://indico.cern.ch/event/313634/

[12] Shiers, "Longterm data preservation", op. cit., 2014.

[13] Shiers, op. cit., 2007.

[14] Shiers, "Hunting the Higgs", op. cit., 2014.

[15] Brumfiel, op. cit., 2011, p.282.

[16] Shiers, op. cit., 2007, p. 219.

[17] Shiers, "Hunting the Higgs", op. cit., 2014.

[18] Shiers, "Longterm data preservation in HEP", op. cit., 2014.

[19] Shiers, "Hunting the Higgs", op. cit., 2014

[20] Shiers, "Hunting the Higgs", op. cit., 2014.

BYTE will need to ensure that the externalities research examines the externalities identified by the WLCG. First, it demonstrates the need for and creation of new technologies and infrastructures to handle the large quantities of data associated with new scientific processes. Furthermore, it also demonstrates that high velocity is a relative concept that depends upon the volume and complexity of the data being processed. Furthermore, the WLCG example also demonstrates that economic and innovation benefits realised through educational provisions (e.g., training future scientists) as well as direct advances in technology development or economic advances in terms of job creation need to be considered. Finally, it also demonstrates that high volume or high velocity data needs to be considered as a long-term project. Not a set of skills and capabilities that can be developed as part of short-term planning.

2.1.2 US Big Data Research and Development Initiative

On the 29 March 2012, President Barak Obama announced a new Big Data Research and Development Initiative; a $200 million programme to improve the "ability to extract knowledge and insights from large and complex collections of digital data" and to "help solve some of the Nation's most pressing challenges".[21] The initiative is being actioned through a range of Data to Knowledge to Action activities by different government agencies. The initiative is of particular interest to BYTE because it aligns with many of the actions and initiatives being undertaken in BYTE. Furthermore, it is characterised by three key features. First, it represents a national, policy initiative that is backed up by a clear action plan and responsible entity. Second, it is a cross-sector, cross-domain initiative that specifically seeks to engage a range of different stakeholders. Finally, it is focused on technological and economic innovation, but it is also cognisant of the social impacts associated with big data.

The White House described the initiative as a plan to leverage big data for three primary aims. These specific aims include:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyse, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use big data technologies.[22]

The US policy initiative seeks to identify and address technological issues, process issues and training and skill development to enable scientists, policy-makers and practitioners to access insights resulting from big data analytics. However, unlike other policy initiatives, this one has a clear action plan and a particular body responsible for ensuring steady progress.

The initiative is being managed by a Big Data Senior Steering Group (BDSSG) within the Networking and Information Technology Research and Development (NITRD) Program. The NITRD program is an umbrella organisation through which Federal agencies coordinate their networking and information technology research and development activities.[23] The BDSSG

[21] Office of Science and Technology Policy, "Obama administration unveils "big data" initiative: Announces $200 million in new R&D investments", *Press release*, 29 March 2012, p. 1. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf
[22] Ibid., p. 1.
[23] Networking and Information Technology Research and Development Program, "Welcome to the NITRD program", no date. https://www.nitrd.gov

identifies areas of interest relevant to big data, and has recently constructed a vision for big data in the US as well as a series of action steps that different agencies can undertake individually and collaboratively to assist in achieving that vision. Specifically, the vision foresees:

> [A] Big Data innovation ecosystem in which the ability to analyze, extract information from, and make decisions and discoveries based upon large, diverse, and real-time data sets enables new capabilities for federal agencies and the nation at large; accelerates the process of scientific discovery and innovation; leads to new fields of research and new areas of inquiry that would otherwise be impossible; educates the next generation of 21st century scientists and engineers; and promotes new economic growth." [24]

This vision specifically views big data as a series of resources and activities that already exist and which need to be steered in ways that benefit the US economy and citizens. Therefore, this initiative is not about developing big data, but about directing it appropriately. The action items associated with this vision reflect this:

- Leverage existing big data foundations, technologies, etc. to propel new technologies and techniques
- Develop mechanisms for understanding the "trustworthiness of data and resulting knowledge" in addition to supporting R&D necessary to undertake data analytics
- Tackle challenges around infrastructure (though access to resources and cyber infrastructure),
- Improve education and training to meet increased demand[25]

In addition to these goals, the vision and action items also mention the creation of cross-sector partnerships and new gateways for cross-agency collaboration as well as ensuring the sustainability of key data assets and resources.[26] While the vision and action items are still the subject of a stakeholder consultation, their conceptualisation and the continuing activities of the BDSSG demonstrate a commitment to progressing the initiative and ensuring appropriate outcomes. This strategy is also being used by the BYTE project, as like the US initiative, it is intended to provide a clear vision and plan of action for achieving that goal, as well as an intention to generate stakeholder commitment to the vision and resulting action steps though stakeholder consultation.

The White House initiative is also characterised by clear cross-sector and cross-agency activities involving a range of different stakeholders. First, the initiative was initially launched with the involvement of six different government agencies and departments working together. In fact, six different departments and agencies were contributing resources to the overall sum of $200 million:

- National Science Foundation
- National Institutes of Health
- Department of Defense
- Department of Energy
- US Geological Survey

[24] Networking and Information Technology R&D Program Big Data Senior Steering Group (BDSSG), "The National Big Data R&D Initiative: *Vision and Actions to be Taken", Predecisional draft,* 26 September 2014. https://www.nitrd.gov/nitrdgroups/images/0/09/Federal_BD_R&D_Thrusts_and_Priority_Themes.pdf
[25] NITRD BDSSG, op. cit., 2014.
[26] Ibid.

- Defense Advanced Research Projects Agency (DARPA)

In some cases, these agencies had formulated plans to work together to greatly improve "tools and techniques" for accessing, organizing and gleaning useful information from available digital data.[27] In others, their commitment to cross-agency activity was represented by their participation in the initiative. In addition to government departments working together, the initiative also foresees big data gains and innovations across a range of different sectors, specifically:

- Health and disease
- Earth Science / environment
- Education
- Biology and genetics
- Defense, and
- Intelligence gathering

In consequence, the benefits and focus areas of the initiative are also cross-disciplinary. Furthermore, the "BDSSG was initially formed to identify big data R&D activities across the Federal Government, offer opportunities for agency coordination, and jointly develop strategies for a national initiative."[28] As such, the coordinating organisation itself is committed to fostering stakeholder partnerships, and it solicits commentary and feedback from multiple groups of big data stakeholders via consultation events. This feeds into the activities being undertaken under the Data to Knowledge to Action initiatives, where the National Science Foundation is committed to "the development and implementation of novel, multi-stakeholder partnerships that promise progress in big data discovery, education and innovation".[29] Like BYTE, the White House policy initiative views the collaboration of multiple, cross-sector groups of stakeholders as paramount in achieving the aims of the activities.

The White House initiative is focused on capturing the economic and social benefits of big data, but it is also framed by an awareness of the potential negative impacts of big data practices. For example, the initiative itself is intended to result in job creation[30] as well as to fuel start-ups and provide broader economic gain[31]. However, alongside this exists a commitment to recognising and analysing the potential pitfalls related to privacy, ethics and social issues. Specifically, one of the data to knowledge to action activities is the creation of a Council for Big Data, Ethics and Society headed by key privacy, ethical and information science experts.[32] Additionally, the White House initiative has also included commissioning a *Big Data and Privacy Report*, released in May 2014 that examined the potentials and pitfalls of big data analytics in both the public and private sectors. While the Council has yet to produce specific material, the *Big Data and Privacy Report* has recommended revising existing legislation, and, in some cases, developing new legislation, to better protect people's

[27] OSTP, op. cit., 2012, p. 1.
[28] NITRD, "Request for Input (RFI)-National Big Data R&D Initiative", 2 October 2014. https://www.nitrd.gov/bigdata/rfi/02102014.aspx
[29] National Science Foundation, "NSF advances national efforts enabling data-driven discovery", *Press Release 13-188*, 12 November 2013. http://www.nsf.gov/news/news_summ.jsp?cntn_id=129244
[30] Wait, Patience, "White House Unveils Big Data Projects, Round Two", *InformationWeek*, 12 November 2013. http://www.informationweek.com/big-data/big-data-analytics/white-house-unveils-big-data-projects-round-two/d/d-id/1112226?
[31] NSF, op. cit., 2013.
[32] White House, "Fact Sheet: Data to Knowledge to Action", 12 November 2013. http://www.whitehouse.gov//sites/default/files/microsites/ostp/Data2Action%20Announcements.pdf

fundamental rights in an age of "big data". This includes protecting consumers, non-US citizens and other "protected groups" to prevent discrimination and irresponsible practices.[33] Nevertheless, both these aspects are overshadowed by the focus on technological innovation and insights. In BYTE both facets of the big data landscape generate equal traction, and the potential positive and negative externalities are considered in tandem. As such, neither the potential benefits, nor the potential negative impacts overshadow the other.

Where the initiative departs from BYTE is a focus on training and skill development, which provides important information that the project can use to enhance its message. The overall aims of the White House initiative as well as the vision developed by the BDSSG place training at the forefront of the activities needed to develop the sector. Some of these training and skill development programmes are designed to generate interest and "buzz" among stakeholders not directly involved. For example, the Department of Defense is using prize competitions to stimulate interest in these areas[34] and the NIH is allowing open access the 1000 genomes project data on the cloud (researchers will only be charged for the computing resources they use) [35].

Therefore, the US Big Data Research and Development Initiative is a clear example of a government seeking to harness big data to effect meaningful societal benefits across a number of sectors, including health science and defense, as well as producing socio-economic advantages for society in terms of prompting skills, training and employment. This initiative is also indicative of the necessity for cross-sector and cross-agency collaboration and implementation that involves a number of different stakeholders in order to achieve its policy aims and objectives. This initiative is also reflective of an increasing trend adopted by national governments to recognise the usefulness of big data for not only research and development, but also utilise the power of big data to foster society-wide policy development.

2.1.3 Australian Public Service Big Data Strategy

In August 2013, the Australian Government released the Australian Public Service Big Data Strategy ("the Strategy"). The Strategy is facilitated by the Australian Information Management Office, a unit of the Department of Finance and De-regulation, and sets out the actions that the Government is taking to harness the opportunities afforded by big data without compromising the privacy of individuals. The Strategy is "Underpinned by six 'big data principles', the strategy aims to position Australia as a world leader in the public sector use of big data analytics to deliver service delivery reform, better public policy and protect citizens' privacy."[36] According to the Strategy, it is intended for Australian government agency senior executives with responsibility for delivering services and developing policy.[37] The strategy is relevant to BYTE as it is an initiative that focuses on the positive externalities of big data, whilst developing ways in which negative externalities can be minimised. The Strategy is supported by a big data specific policy and also provides clear action plans to be carried out within a nominated time frame. Furthermore, it promotes cross-agency and cross-

[33] White House, *Big Data: Seizing Opportunities, Preserving Values*, May 2014. http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
[34] OSTP, op. cit., 2012.
[35] OSTP, op. cit., 2012, p. 3.
[36] Lohman, Tim, "Australia Wants to be World Leader in Big Data Analytics", *Zdnet,* 5 August 2013. http://www.zdnet.com/australia-wants-to-be-world-leader-in-big-data-analytics-7000018963/
[37] The Strategy does not apply to the intelligence, and law enforcement sectors. Commonwealth of Australia, *Australian Pubic Service Big Data Strategy*, 2013, p.5. http://www.finance.gov.au/sites/default/files/Big-Data-Strategy_0.pdf

sector collaboration and development that involves a number of different stakeholders. The Strategy focuses on the analysis of a variety of large data sets produced across a variety of sectors. Overall, the Strategy seeks to improve understanding through enhanced data analytics capabilities - a targeted action that is becoming more widely adopted by public and private sector organisations. The Strategy raises the issues of the value of data held by Commonwealth agencies and the responsibility to realise this value to benefit the Australian public, as well as the need to negotiate the privacy risks of linking, sharing and providing broader access to data. Similarly with BYTE, the Strategy is interested in capturing the socio-economic benefits that flow from big data analytics, as well as promoting technological developments and innovative means of boosting the digital economy.

The Strategy

As previously observed, the strategy aligns with BYTE as it ultimately focuses on capturing the benefits of big data, whilst minimising the negative externalities such as privacy implications. This is encapsulated in the vision of the Strategy:

> The Australian Government will use big data analytics to enhance services, deliver new services and provide better policy advice, while incorporating best practice privacy protections and leveraging existing ICT investments. The Australian Government will be a world leader in the use of big data analytics to drive efficiency, collaboration and innovation in the public sector.[38]

This vision aims to support enhanced services, new services and business partnership opportunities, improved policy development, and the protection of personal information privacy, and leveraging the Government's investments in ICT technologies.[39]

The development of the Strategy was initially identified in the Australian Public Service Information Communication Technology Strategy 2012-2015.[40] The aims of the Strategy are broadly identified as:

- Delivering better services through building capability and improving services;
- Improving the efficiency of government operations by investing optimally and encouraging innovation; and
- Engaging openly through creating knowledge and collaborating effectively.

Meeting these objectives depends primarily on the efficient and effective use and analysis of big data, which will be achieved by adherence to the following six principles of the Big Data Strategy:

- Data is a national asset (to be shared by all);
- Privacy by design;
- Data integrity and the transparency of processes;
- Skills, resources and capabilities will be shared;
- Collaboration with industry and academia; and
- Enhancing open data.

[38] Commonwealth of Australia, op. cit., 2013, p.19.

[39] Ibid., p.20.

[40] Department of Finance and Deregulation, Australian Public Service Information and Communications Technology Strategy 2012-2015, http://agimo.gov.au/ict_strategy_2012_2015/. That Strategy also focussed on developing technologies for analysis and cross department collaboration: Commonwealth of Australia, *Australian Public Service Information and Communications Technology Strategy 2012-2015, p.7*. http://www.finance.gov.au/publications/ict_strategy_2012_2015/.

The six principles of the Big Data Strategy are expected to assist the government to realise substantial productivity and innovation gains as well as help tackle intractable policy and business challenges. The Strategy sets a timeline for concrete actions to put the six principles into action. These actions began with the development of a big data better practice guidance in March 2014, followed by a report on the barriers to big data analytics in July 2014 (discussed below). These actions will be followed by a push to ensure that the ICT industry and education sectors can supply the skills necessary for big data analysis, develop a guide to responsible data analytics. The focus on skills and the development of a guide for responsible big data analytics will be useful reference documents to compare to similar strategies undertaken by the BYTE case study subjects. Moreover, the Strategy proposes two ongoing projects: developing an information asset register; and monitoring technical advances in big data analytics.[41]

Finally, the Strategy was supplemented by the *Better Practice Guide for Big Data*[42] in April 2014, which gives guidance on establishing a business requirement for a big data capability, implementation, information management and big data project management. Establishing a business requirement for big data is premised on the standard considerations of cost and return on investment, but also on an agency's current and future:

- Strategic objectives
- Business model
- Data availability
- (Maturity of) technology and capability
- Availability of skilled personnel to manage data acquisition and analysis.

The relationship between business and government and big data analytics for societal gain is also recognised by BYTE.

The data

The Strategy will harness the vast amount of data already held by the Australian government and its agencies. The government produces data as a result of its administrative and policy development activities and interactions with the Australian public. Big data exists in structured, semi-structured and unstructured forms, including data generated by machines such as sensors, machine logs, mobile devices, GPS signals, as well as transactional records. According to IBM, some of the most common types of data being used in big data analytics include internal transactional, log and event data.[43] The volume of data has grown due to the adoption of new technologies and the production of an increasing amount of structured and unstructured data outside of government. It is accepted that "the digital economy has seen an exponential increase in the production of data, not least in government-collected data about citizens and businesses, organisations' internal operations and its own interactions with external parties such as suppliers and communities."[44] In that regard, Australian Government agencies alone have installed an additional 93,000 terabytes of storage during the period

[41] Lohman, op.cit.,2013.

[42] Australian Public Service, *Better Practice Guide for Big Data*, Australian Government, April 2014. http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf

[43] Commonwealth of Australia, op cit., 2013, p.5.

[44] Wedutenko, Alexandra and Lisa Keeling, "Australia: Big Data and the Public Sector: Strategy and Guidance", *Mondaq,* 3 June 2014.
http://www.mondaq.com/australia/x/317214/Constitutional+Administrative+Law/Big+data+and+the+public+sector+strategy+and+guidance

2008-2012[45] to cope with increasing data production. Thus, the Strategy involves big data in the sense that it meets that the relevant definitions of big data identified in Deliverable 1.1 of BYTE. In fact, the meaning of big data analytics adopted by the strategy refers explicitly to the volume, velocity and variety of the big data at the centre of the Strategy:

1. The data analysis being undertaken uses a high volume of data from a variety of sources including structured, semi-structured, unstructured or even incomplete data; and
2. The size (volume) of the data sets within the data analysis and velocity with which they need to be analysed has outpaced the current abilities of standard business intelligence tools and methods of analysis.[46]

The Strategy means that the proposed analytics of this big data can increase the value of the asset to the government and to Australian society.[47] However, whilst the Strategy recognises that the magnitude and nature of the value of the data varies depending upon the industry sector, the Strategy anticipates substantial productivity and innovation gains from the use of all the data.[48] Predicted gains are expected in terms of leading to better service delivery and supporting government agencies in more efficiently carrying out their duties.

Therefore, big data analytics promoted by the Strategy reiterates the Australian government's dynamic approach to big data analytics in the real time as the most efficient means to produce benefits. "In traditional data analysis, structured sets of data were analysed often using Structured Query Language (SQL). While SQL may still be used for particular purposes, a feature of big data analytics is that all of the data, including structured, unstructured and messy data, is analysed in real time."[49]

The benefits

The Australian Government foresees that the Strategy will produce a number of society-wide benefits for the government and the Australian people:

> Big data analytics can be used to streamline service delivery, create opportunities for innovation, and identify new service and policy approaches as well as supporting the effective delivery of existing programs across a broad range of government operations - from the maintenance of our national infrastructure, through the enhanced delivery of health services, to reduced response times for emergency personnel.[50]

These benefits are expected to be realised especially in the area of public policy and service delivery. In that regard, the Strategy promotes big data analytics to achieve enhanced services by providing better information about service delivery outcomes and inform future models for the provisions of these services as well as identifying where gaps exist under current service delivery arrangements. The Strategy also anticipates that big data analytics will enable government agencies to better target services to those that need them, thus allowing more efficient and effective delivery of services, as well as enabling agencies to improve services by tailoring service delivery based on the individual needs of businesses and communities.

[45] Department of Finance and Deregulation. Australian Government ICT Expenditure Report 2008-09 to 2011-12. http://agimo.gov.au/files/2012/04/Australian-Government-ICT-Expenditure-Report-2008-09-to-2011-12.pd
[46] Commonwealth of Australia, op. cit., 2013, p. 8.
[47] Ibid., p.5.
[48] Ibid., p.13.
[49] Wedutenko and Keeling, op. cit., 2014.
[50] Ibid., p.5.

Additional benefits are expected in terms of new services and business partnerships. The industry, research and academic sectors have been working on big data analytics projects for some time and continue to invest heavily in the skills, technologies and techniques involved with big data analysis. These sectors are also identified as being key custodians of valuable data collections, and potential partners with the Government, for the delivery of insights from big data analytics that promote the public good.[51] Government will work with these sectors by leveraging and sharing expertise in big data analytics and related fields, and will also work with these sectors to promote the continued development of skills in this area. The government is also strengthening the skills across agencies through initiatives such as the Data Analytics Centre of Excellence. This purpose of this initiative is to bring together representatives from across government and from a multitude of disciplines to share technical knowledge, skills and tools whilst building analytics capability. The development of new services based on insights derived from the analytics process, as well as industry developments and the maturity of tools and services that utilise big data analytics that will create entirely new business opportunities and industries based on using open government data. It follows that insights into business problems will also be had as a result of the Strategy. Further, unanticipated correlations and discoveries may provide important insights that could lead to innovative solutions that might not otherwise have be reached. These insights can provide opportunities to act and respond more rapidly to information and trends as they occur.[52] BYTE places an emphasis on the development of partnerships and cross-sector collaborations that involve a number of stakeholders. The Strategy clearly promotes the same in the pursuit of capturing the full value of big data for society.

Additional benefits include policy and economic gains. For example, one outcome of the Strategy is to utilise the results of data analytics to support better policy development by strengthening evidence-based decision-making and provide more immediate information about policy settings and their impacts. An additional significant benefit expected from the Big Data Strategy is the expected creation of new ICT jobs and potentially new professions, particularly as there is a major shortage of data scientists with experience in big data analytics.[53] This is supported by Gartner research that estimates, by 2015, big data demand will reach 4.4 million jobs globally, with two thirds of these positions remaining unfilled.[54] There are also related initiatives in terms of addressing the workforce challenge. The update to the National Digital Economy Strategy outlines initiatives for the completion of the development of a new curriculum for technologies, and the promotion of careers in ICT to school students. Other related initiatives are aimed at boosting the workforce in this area to further promote and support the development of skills and interest in data mash-ups, apps and visualisations, all of which are central to big data analysis. One such example is including GovHack.[55] GovHack is a 48-hour, competitive event that encourages teams to find new ways to produce innovative solutions with open data.

Overall, the Strategy is an example of why big data analytics is of increasing importance for big data actors of the future: "The benefits of big data analytics are no secret; government recognises that their use can improve decision-making, targeting and delivery of services, and

[51] Commonwealth of Australia, op.cit., 2013, p.18.
[52] Ibid., p.13.
[53] Ibid., p.17.
[54] Gartner, "Gartner Reveals Top Predictions for IT Organisations and Users for 2013 and Beyond", 24 October 2012. http://www.gartner.com/it/page.jsp?id=2211115
[55] GovHack, "GovHack", 2014. www.govhack.org

thus productivity, which, in turn, can substantially reduce government administrative costs."[56] Nevertheless, the Strategy recognises that promoting big data analytics will also engender a number of challenges.

Challenges

Capturing the benefits to be produced by this big data initiative presents challenges, similar to those of interest to the BYTE project, including the legal and ethical implications of privacy. Insight into the potential of the Big Data Strategy reveals, "The challenge lies not in convincing agencies to use big data but in actually analysing it effectively and lawfully."[57] Commentators also recognise, "Advances in technologies, including cloud computing, will make big data analytics more technologically accessible for Agencies, but may also increase the associated risks of breach of confidentiality, privacy and security."[58] In order to assess the security of the infrastructure, an Agency contemplating big data analytics needs to ensure scalability of their infrastructure to ensure the infrastructure is optimised for very fast capture and retrieval, which means understanding the likely size of the data it will capture and store.[59]

Protection of privacy during the implementation of the Big Data Strategy also necessitates developments such as incorporating "privacy by design" into big data analytics projects, and proactively ensuring the privacy of the individual's data and information. These strategies signal the adoption of better practice methodologies that address the potential risk to privacy posed by big data analytics and "the mosaic effect". Specifically, it will prevent practitioners from combining disparate data sets to construct detailed and identifiable information about specific individuals contained within those datasets.[60]

The Department of Finance and De-Regulation has indicated that the Government intends to own the intellectual property rights in new databases developed as part of its big data analytics as a way of dealing with any future contention over such legal issues raised by the Strategy. However, attribution of respondents has been identified as playing a vital role in the compliance with local copyright laws.[61] Therefore, the typical legal hurdle presented by big data initiatives such as the Big Data Analytics Strategy are addressed as part of the policy backing the Strategy which promotes proactive explorations in this area. The policy can also act as an example for other similar strategies in Australia and overseas.

The Strategy discussed here is useful for BYTE as it presents challenges to those wishing to effectively and compliantly capture the true value of big data for society. The symbiotic nature of benefits and challenges produced by the strategy are aptly described: "Effective use of big data has the capacity to significantly improve Government service delivery, operations and policy development, but there are risks associated with big data analytics that will require careful consideration at all stages of any big data analytics exercise."[62] This demonstrates that the Australian Strategy, like BYTE, is a strong policy backed initiative that aims capture the benefits of big data, particularly through analytics, whilst providing for concrete means to

[56] Wedutenko and Keeling, op. cit., 2014.
[57] Ibid.
[58] cited in Wedutenko and Keeling, op. cit., 2014.
[59] Ibid.
[60] Breeden, II, Josh, "Worried about security? Beware the mosaic effect", *GCN: Technologies, Tools and Tactics for Public Sector IT*, 14 May 2014. http://gcn.com/articles/2014/05/14/fose-mosaic-effect.aspx
[61] cited in Wedutenko and Keeling, op. cit., 2014.
[62] Ibid.

reduce negative externalities, including technological requirements for the protection of personal information privacy and compliance with other relevant laws.

2.1.4 United Nations Global Pulse Initiative

The UN Global Pulse initiative ("Global Pulse"), heavily funded by the governments of Australia, Sweden and the UK, is an effort to harness the power of big data and analytics in order to better understand how the world is changing and how communities are impacted by changes. Thus, the initiative is about maximising the information and insight gained through big data analytics for humanitarian causes, and it raises a number of positive outcomes of big data that are relevant to BYTE. Further, because Global Pulse is essentially an information initiative that matches interested UN agencies, governments and partners with private sector organisations that have the capabilities required to investigate research questions and develop working proofs of concept, it represents a positive example of cross-sector and cross-industry collaboration. Global Pulse works with partners to design experiments, coordinate research, evaluate results and communicate findings. Global Pulse supports UN System partners in pursuing the institutional adoption of tools and approaches that have been proven effective through R & D. The initiative is supported by the Global Pulse "innovation laboratory", which aims to increase access to real-time data advising UN policy development. The recently released a whitepaper[63] refers to the initiative as a "critical milestone" to harnessing information related to a country's economic or social state. In that regard, Robert Orr, Assistant Secretary-General to the UN's policy coordination and strategic planning observes, "it's no longer any question of whether or when data science will be applicable to the work of the United Nations. It's about how."[64] Global Pulse is described as:

> A flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action. The initiative was established based on recognition that digital data offers the opportunity to gain a better understanding of changes in human well being, and to get real-time feedback on how well policy responses are working.[65]

This initiative ties in with a number of interrelated aspects of the BYTE project that examines how positive externalities of big data can be captured to produce benefits for local and global communities, particularly (as previously mentioned) it aims to do so through a variety of cross-sector collaborations and partnerships. The objectives of Global Pulse also highlight the great value (e.g. societal and economic benefits) to be gained through the collection and use of big data, particularly through big data analytics. The Global Pulse initiative pursues the following main objectives:

1. Increasing the number of Big Data for Development (BD4D) innovation success cases;
2. Lowering systemic barriers to big data for development adoption and scaling; and
3. Strengthening cooperation within the big data for development ecosystem[66]

---

[63] UN Global Pulse, *White Paper*: *Big Data for Development: Opportunities & Challenges*, May 2012. http://www.unglobalpulse.org/projects/BigDataforDevelopment

[64] Hilvet, John, "Australian Funds UN Big data Initiative", *IT News for Australian Businesses,* July 2012. http://www.itnews.com.au/News/308301,australia-funds-un-big-data-initiative.aspx#ixzz3FKgYFC8N

[65] United Nations Global Pulse, "About", no date. http://www.unglobalpulse.org/about-new

[66] Ibid.

Thus, a main focus of this initiative is on utilising big data technologies in novel ways to produce socio-economic benefits. Global Pulse uses data from social networks, blogs, cell phones and online commerce to "transform economic development and humanitarian aid in poorer nations."[67] The efforts by Global Pulse and a growing collection of scientists at universities, companies and non-profit groups have been given the label "Big Data for development." [68] The goal is to bring real-time monitoring and prediction to development and aid programs. Projects and policies, they say, can move faster, adapt to changing circumstances and be more effective, helping to lift more communities out of poverty and even save lives.[69] Achieving these objectives are dependent on the quantity, quality and accuracy of the data used, as well as ensuring that access to these kinds of data continue to be available.

The data

Global Pulse is drawing on a number of new data sources, including data that can be grouped under umbrella terms such as data exhaust, online information, physical sensors and citizen reporting or crowd-sourced data. These data types are understood to mean the following:

- Data exhaust - data that is passively collected transactional data from people's use of digital services like mobile phones (call detail records, location data, airtime purchase patterns), making purchases, transferring remittances or mobile money, etc., and or operational metrics and other real time data collected by UN agencies, NGOs and other aid organisations to monitor their projects and programmes (e.g. stock levels, school attendance etc. All of these digital services create networked sensors of human behaviour;
- Online Information – web content such as news media and social media interactions (e.g. blogs, Twitter), web searches, news articles obituaries, e-commerce, job postings; this approach considers web usage and content as a sensor of human intent, sentiments, perceptions, and want;
- Physical Sensors – satellite or infrared imagery of changing landscapes, traffic patterns, light emissions, urban development and topographic changes, etc.; this approach focuses on remote sensing of changes in human activity; and
- Citizen Reporting or Crowd-sourced Data – Information actively produced or submitted by citizens through mobile phone-based surveys, hotlines, user-generated maps, etc. While not passively produced, this is a key information source for verification and feedback.[70]

However, the availability and types of digital data will indeed differ from country to country:

> Countries with high mobile phone and Internet penetration rates will produce more data that is directly generated by citizens. Countries with large aid communities will produce more programme-related data than less aid-dependent countries. Countries with vibrant local business environment will offer greater opportunities for private sector involvement.[71]

[67] Lohr, Steve, "Searching Big Data for Digital Smoke Signals", *New York Times,* 7 August 2013. http://www.nytimes.com/2013/08/08/technology/development-groups-tap-big-data-to-direct-humanitarian-aid.html?_r=1&

[68] Ibid.

[69] Lohr, op. cit., 2013.

[70] United Nations Global Pulse, op. cit., no date.

[71] Ibid.

Further, the availability of data also varies between age groups, economic income brackets, gender and geographic location. However, Global Pulse research projects endeavour to address such biases. The variance in data and data means that Global Pulse has developed means and partnerships for exchange off and access, to the data necessary to meet the objectives and aims of Global Pulse. Thus, Global Pulse is an example of a partnership in terms of accessing data from outside the UN, such as national statistical data and accessing data from international agencies. Whilst the majority of the requisite data are already available in the public domain, Global Pulse also is also challenged to develop tools to both capture and responsibly use the raw data. This is especially relevant when personal data is gleaned from social networks, which are a favoured source of data because it enables Global Pulse to see big data through a human lens. This challenge has driven Global Pulse to develop new and innovative ways of overcoming barriers to data access, such as encouraging cross-sector collaboration, especially in terms of getting private organisations to open up their data stores, that were previously kept confidential and viewed as an asset for competitive advantage only. This example of cross-sector collaboration provides a clear example of the importance of big data partnerships and collaborations, an issue of interest to BYTE.

Overcoming barriers to data access

The implementation of Global Pulse is an example of where negative externalities, such as privacy implications or intellectual property rights, can present barriers to access to data. One of the principal aims of the BYTE project is to identify ways in which these challenges can be addressed in order to minimise their negative effect. In doing so, companies and organisations will be better placed to responsibly capture the full potential of big data. The Global Pulse initiative presents clear examples of how these barriers manifest.

Global Pulse initiative is committed to the responsible use of big data and usage that accords with human rights. For example, Global Pulse recognises that the risk of re-identification through misuse of big data could lead to discrimination. In consequence, the initiative depends on large data sets of anonymised, aggregated data that can give a sense of how whole populations or communities are coping with shocks that can result in widespread behavioural changes.[72]

Robert Kirkpatrick, Director of Global Pulse, observes that another barrier to accessing more data are that corporate entities often protect data to preserve consumer privacy as well as corporate intellectual property and commercial competitiveness.[73] Global Pulse raises privacy and data protection issues especially because it relies on data from cell phones, which are mobile sensors of human behaviour and provide particularly useful data for development programs, despite these privacy implications. Furthermore, another barrier to accessing more data is created by limited capabilities of finding data signals, especially as valuable data are held by private companies, including mobile phone operators, whose networks carry text messages, digital-cash transactions and location data. So, persuading telecommunications operators, and the governments that regulate and sometimes own them, to release some of the data is a top task for the group. In order to overcome these barriers, Kirkpatrick advocates the concept of "data philanthropy" and the creation of a public "data commons," in which companies contribute large customer data sets, stripped of personally identifying information, for research on development and public health. Kirkpatrick insists that it be a matter of self-

[72] United Nations Global Pulse, op. cit., no date.

[73] Kirkpatrick, Robert, "UN Global Pulse", *Strata Summit presentation,* 2011. *http://www.youtube.com/watch?v=kUCQ8zYsYNQ*

interest, since economically healthy communities are more attractive markets.[74] Data philanthropy describes the exchange of data for the results of the analyses that would benefit companies, such as financial forecasts for organisations and companies. [75] The benefit of data philanthropy is that it can support a company's risk strategy by providing analyses that has been undertaken in real time, but provide businesses with a broader perspective, such as reporting on social crises that may ultimately lead to financial issues that may affect the business granting access to their data. Overcoming these barriers to access are integral to ensuring that the big data technologies and practices employed by Global Pulse have sufficiently varied data to analyse and process.

The technology

Global Pulse initiative fosters projects that utilise a number of big data technologies and tools. Data mining tools, tools for real time analytics and data storage, computing and data visualisation assist data scientists to tackle research questions are vital aspects of this initiative and align with the focus on big data specific technologies and practices that are addressed throughout the BYTE project. The probability of an effective initiative such as Global Pulse is thus dependent upon continued innovation and continuing development of such technologies by public and private sector organisations. This is because technologies employed by Global Pulse enable tracking of diseases and other social and health concerns in real time. For example, Kirkpatrick states, "We're trying to track unemployment and disease as if it were a brand."[76]

Moreover, the big data technologies and practices also result in the creation of relationships across sectors, and within private and public sectors:

> In Indonesia, for example, Global Pulse has worked with both Crimson Hexagon, a start-up, and SAS Institute, a large data analytics software company, to mine Twitter messages and other online media for clues to price trends. The smart algorithms must identify not just words, but context and often sentiment. "I had rice for breakfast" is not a signal. "The price of rice is getting scary" is. The research found that surges in online mentions accurately capture price increases a month or two before official statistics.[77]

Thus, Global Pulse is a big data initiative that recognises the value of cross-sector collaboration to produce optimum results from the big data technology process employed, as well as ensuring access to the technology that is required to produce meaningful relationships from the data. These relationships, and the necessity for specific technology and practices, have also promoted technological innovation. For example, Global Pulse has developed an open source tool called "HunchWorks". This was created out of the necessity to address issues as they arise and not only after there is sufficient hard evidence. Kirkpatrick describes the reason for this being that more often than not, experts will have a hunch that a trend is developing or something is occurring but they do not wish to openly publish their hunch if they do not believe they can substantiate it with concrete evidence. The problem with this is that by the time the evidence is available, it is often too late to respond effectively to mitigate any damage caused by issues or crises or indeed prevent something from occurring. HunchWorks enables users to upload their hunch with some evidence to be viewed by

[74] cited in Lohr, op. cit., 2013.
[75] Kirkpatrick, op. cit., 2011.
[76] cited in Lohr, op. cit., 2013.
[77] Lohr, op. cit., 2013.

nominated users only. These users can respond to support the hunch or refute it with evidence further evidence. The effect of this is the potential for even earlier detection of changes in societal and personal circumstances. This ultimately enables for more timely responses to social challenges and crises as they are developing.

The projects

Global Pulse Initiative has produced a number of projects with a central humanitarian focus. Five general "proof of concept" projects were completed in 2005, which are said to demonstrate:

> Feasibility and utility of real-time digital data (such as social media, online news, mobile phone surveys, and online food prices) to answer questions relevant to decision-makers (such as unemployment trends, coping strategies, food prices, and public perceptions).[78]

More recently, projects between 2012 and 2014 apply new methodologies and technologies towards specific programme and policy related questions. Projects are identified in-country by UN agencies and their local partners, and the public sector. These are carried out in Pulse labs which are physical centres of innovation and R&D that bring together government, the UN, and local partners in academia and the private sector to test, refine and scale methods for using digital data streams to support development goals.[79] Pulse labs are a crucial development of Global Pulse and bring together experts from government, NGOs and private companies in the region to work on applying innovative analytic approaches and data science to thematic social development challenges. Approaches for exploration adopted within the labs include: social media and twitter analysis; mobile phone data analysis; rapid mobile surveys; and geo-spatial mapping. Pulse labs are located in New York, Jakarta, Indonesia and Uganda. The research labs are initially working on demonstration projects to show the potential of the technology. William Hoffman, an Associate Director who leads the data-driven development programme at the World Economic Forum, a forum that has partnered with global pulse, opines, "The larger role of Global Pulse is as a catalyst to foster a data ecosystem for development, bringing together the private sector, universities and governments".[80]

Overall, Global Pulse runs innovation projects in which it partners with organisations that have access to relevant sources of big data, data analytics technologies, and data science expertise, as well as with UN agency and government ministry "problem owners" grapping with challenges that could benefit from new insights and real-time measurement tools, to discover, build and test high-potential applications of big data. Its innovation programmes focus on sectors such as food security, agriculture, employment, infectious disease, urbanisation, and disaster response, as well as cross-cutting issues such as M&E and privacy protection.[81] These benefits produced by global Pulse are discussed below.

Capturing the benefits

Global Pulse's use of big data has proven to be beneficial as a humanitarian effort in its own right and in assisting related efforts. This is because Global Pulse was launched with the primary aim of using best practices in the big data industry to make faster and better-informed

[78] United Nations Global Pulse, op. cit., no date.
[79] Ibid.
[80] cited in Lohr, op. cit., 2013.
[81] "United Nations Global Pulse", *Wikipedia,* no date.
http://en.wikipedia.org/wiki/United_Nations_Global_Pulse

responses to humanitarian crises.[82] For example, a project using social media analysis to track public concerns in Indonesia and the US[83] has already showed the value such methods hold for crisis response and social science research as a whole[84]. Further, research by Global Pulse has found that analysing Twitter messages can give an early warning of a spike in unemployment, price rises and disease. Kirkpatrick attests that such "digital smoke signals of distress" usually come months before official statistics — and in many developing countries today, there are no reliable statistics.[85] When such data is aggregated, it can indicate how people cope in times of stress. Some examples of beneficial insights gained through the Global Pulse initiative include:

- The way in which people add money to recharge their mobile phone accounts can indicate economic standing and circumstances;
- Online food prices can signal prices spikes and food shortages in their early stages;
- Publicly shared information on social networks can reveal topics of concern, and how people feel about job prospects or their future; and
- Analysis of Internet searches can help public health official detect outbreaks of diseases such as the flu or cholera faster than ever before.[86]

Once digital signals are anonymised and analysed, they reveal information about changes and the implications of changes in real time. This is in stark contrast to the previous situations where analysts relied on government-collected data such as surveys and census that could provide a similar result, but long after the change, the implication and the resolution had occurred. This initiative is a useful example for BYTE of when real time analysis of social, economic and health concerns enable efficient global responses. This is especially so when we consider this initiative's focus on how digital signals can be used to strengthen internal development, and responses to social, economic or political problems can be implemented around the time they needed. This turns information into solutions, rather than information in retrospect. Nevertheless, further development of Global Pulse is somewhat dependent upon partnerships with private sector companies who hold a wealth of data to about their customer preferences.

Overall, Global Pulse provides a meaningful big data story with a focus on collection and use of big data to support the detection of, and timely response to, humanitarian issues and crises. This is also an important story for BYTE as it promotes innovative means of securing data through cross-sector collaborations through the development and implementation of the concept of data philanthropy, a useful tool to promote open access to, sharing and exchange of large data sets.

2.1.5 eBay Inc. Big Data Analytics Programme

eBay Inc. ("eBay") implements a big data analytics program to add value by machine learning, data mining, economics, user behaviour analytics, information retrieval and visualisation. The eBay big data analytics program is a high-functioning system implemented

[82] Burn-Murdoch, John, "Big Data: What Is It and How Can It Help?", *The Guardian Data Blog,* 27 October 2012. http://www.theguardian.com/news/datablog/2012/oct/26/big-data-what-is-it-examples
[83] Lopez, Giselle and Wayne St. Amand, "Discovering Global Socio Economic Trends Hidden in Big Data", *UN Global Pulse Blog,* 15 July 2012. http://www.unglobalpulse.org/discoveringtrendsinbigdata-CHguestpost
[84] Burn-Murdoch, op. cit., 2012.
[85] Lohr, op. cit., 2013.
[86] United Nations Global Pulse, "An Animated Introduction to the UN's Global Pulse Initiative", *Youtube,* 2012. http://www.youtube.com/watch?v=kYg80Op2whM

for data collection, real time analysis, targeted distribution, and automated response etc. that supports the business aim of collecting data relating to "detail-by-detail, minute-by-minute"[87] from its users. The wide variety of data eBay works with includes user data, user behaviour data, transaction data, items data, feedback data, and search query data. The deluge of data is helping eBay to emulate the know-how that customers used to get from a local shop owner; the only difference is it is trying to achieve this across its global auction sites.[88] With more than 100 million active users globally,[89] the programme is the source of useful and meaningful information that assists business units across the company. It follows that the programme produces a variety of benefits to consumers, in addition to the overall enhanced business model and added value for eBay.

The purpose of the programme was articulated by David Stephenson, head of global business analytics at eBay when speaking at a recent Gartner CRM Summit in London, when he said "the auction site's goal is to make shopping successful. As a marketplace, eBay's primary business involves being successful from a buyer's and a seller's perspective. The company is using analytics to help it understand its customers better." [90] Stephenson further describes the programme ambition as taking the kind of personalisation possible in a small shop and apply it to the world of eBay: "In a small store, engaging the customer is key, helping them with search and recommendations, understanding their preferences and learning from existing customers."[91] Consequently, eBay's big data analytics program implements big data technologies and practices in the pursuit of an improved business model that produces commercial gain. This big data story aligns with BYTE as an example of the growing importance of commercial entities adopting big data strategies in order to capture the positive externalities of big data to grow their business. This big data story is also useful to BYTE, as the eBay big data analytics programme is moving towards more open source technologies and is interested in the cross-sector sharing of data.

However, the programme is not without a number of challenges that arise in relation to web analytics. Stephenson observes,

> Web analytics is like having a video camera mounted on the head of every customer going into a supermarket. Recording everything every customer does generates 100 million hours of customer interaction [per month], creating an unmanageable amount of customer data. We need to understand customers, learn from our customers and apply data science techniques to allow us to get more data and new types of data.[92]

However, despite the challenges presented by the big data analytics programme, its existence was necessitated out of the need to remain competitive by extracting the full value of the data that would otherwise have been passively held by eBay.

---

[87] cited in Arthur, Lisa, "The Surprising Way eBay used Big Data Analytics to Save Millions", *Forbes,* 23 August 2012. http://www.forbes.com/sites/lisaarthur/2012/08/23/the-surprising-way-ebay-used-big-data-analytics-to-save-millions/

[88] Saran, Cliff, "Case Study: How Big Data Powers the eBay Customer Journey", *computerweekly.com,* 29 April 2014. http://www.computerweekly.com/news/2240219736/Case-Study-How-big-data-powers-the-eBay-customer-journey

[89] Arthur, op. cit., 2012.

[90] cited in Saran, op. cit., 2014.

[91] Saran, op. cit., 2014.

[92] Ibid.

The data and supporting technologies

eBay reportedly receives an estimated 100 terabytes of new data every day.[93] In 2013, eBay Director of Data and Data Infrastructure, Alex Liang, informed delegates of the Teradata Big Data Analytics summit that the website has more than 50,000 product categories with more than US$3500 goods sold every second. The fact that almost everybody accessing eBay listings is doing so through the use of a smart phone means that Liang's team has access to more and more data, which requires processing. Liang specifies the reason for big data analytics: "For eBay, data is about value so if you cannot get value from big data you should not even work on it."[94] He further provides, "We are facing very aggressive competition from other sites so data is the biggest advantage for eBay. Every business initiative we make is based on data."[95] eBay uses several algorithms to determine potential improvements in the overall user experience.[96] At Big Data World Europe 2012, Dr Neal Sundersan, Senior Director and Head of eBay Research Labs, discussed how his team conducts a variety of activities such as machine learning, data mining, economics, user behaviour analytics, information retrieval and visualisation. These practices are conducted in relation to a wide variety of data, including user, users behaviour, transaction, items, feedback and searches.[97] However, a big data challenge is produced by data associated with these processes, but this has led to a number of technological innovations within the company. Despite data collection of this magnitude, eBay's 1200 internal business intelligence users faced problems getting the value from it which is why eBay has developed a big data program. Internal data users, who range from data scientists to sales directors, require regular reports. In 2011, eBay began rolling out three different platforms, all of which support a particular type of analytics crucial to eBay's business units:

- Enterprise Data Warehouse platform is used for corporate BI reporting, provided by Teradata;
- Singularity is a 40 petabyte Discovery platform for website behavioural analytics. "This allows eBay to test ideas on the site and assess what works, such as testing whether site visitors prefer bigger pictures in search results"[98]; and
- Hadoop is a 40 petabyte Hadoop cluster for technical analytics such as counterfeit detection and image classification.[99]

By implementing these technologies and practices, eBay turns it analytical lens on online behaviours and website traffic patterns, as well as its core business element, that being IT infrastructure. The results of big data analytics of the IT infrastructure revealed that a number of underutilised servers, misconfigured devices and other inefficiencies with the result of

[93] Barwick, Hamish, "E-commerce Company Uses Three Business Intelligence Platforms to Support Analytics", *CIO,* 9 May 2013. http://www.cio.com.au/article/461364/ebay_bids_big_data_challenge_/

[94] cited in Ibid.

[95] cited in Barwick, op. cit., 2013.

[96] Booth, Corey, Sesh Iye and Fabrice, "Technology Trends: Deciding Which Ones Will matter", *BCG Perspectives*, 24 March 2011.
https://www.bcgperspectives.com/content/articles/information_technology_information_technology_strategy_technology_trends_deciding_which_ones_will_matter/

[97] Data Science Series, "How eBay Performs Big Data Research to Create New Insights for its Business", *Data Science Stories,* 2012. http://datascienceseries.com/stories/how-ebay-performs-big-data-research-to-create-new-insights-for-its-business

[98] Saran, op. cit., 2014.

[99] Barwick, op. cit., 2013.

these findings, eBay was able to repurpose thousands of servers and save millions in capital expenditure within the first year.[100]

Therefore, eBay provides a good example of how big data stores become of increasing value to companies that are willing to implement big data specific technology and practices as part of an overarching business strategy. Once implemented, a big data analytics programme produces an array of benefits that can assist a company maintaining its competitive advantage.

Benefits to the business

The eBay big data analytics program sought to capture the value of big data held by eBay and ultimately produce a more effective business model. Oliver Ratzesberger, eBay's former Senior Director of Architecture & Operations describes how the big data process produces commercial value for eBay: "(We saw) patterns that were not obvious to the individual technician, ones that only got visibility once you took all of the corporate data and looked at it."[101] The result of this process was "millions of dollars for us in terms of capital savings – just through applying analytics in an area that at first we never thought about using analytics for."[102] Another positive outcome of the big data analytics programme is that it promotes intra-business collaboration. Liang observes, "Because the business environment is much more complex, you cannot have one analyst working independently. People must be working with each other to get deep data insight."[103]

Additional benefits are described in terms of their relationship with the big data technologies and processes employed by eBay. For example, eBay utilises search data because insight into search data allows its users to broaden or narrow searches, leads buyers to related products and optimises the overall experience of eBay.[104] The essence of how eBay derives benefits from its big data analytics programme is set out as follows:

> In essence, what eBay does is use intelligence from advanced users and apply that to help what they call 'The naïve user' (a user who's not good with queries). A lot of effort goes into the first step of cleaning data. De-duplicating user-associated data provides better suggestions for related searches. After that, eBay goes six years back in time to analyse user behaviour. And it does this pretty much in real time."[105]

Another example of eBay's research involving big data is found in the hundreds of thousands of economic experiments its users run every day. Users carry out their own research to be examined such as altering their listings to achieve faster selling times or gain other advantages by things such as offering free shipping, uploading more than two images, offering on a Thursday rather than a Friday. eBay does not have to conduct these economic experiments itself. Its users are doing this continuously and on a large scale. "All that remains for eBay to do is gather the data and put it to good use."[106]

---

[100] Arthur, Lisa, "The Surprising Way eBay used Big Data Analytics to Save Millions", *Forbes,* 23 August 2012. http://www.forbes.com/sites/lisaarthur/2012/08/23/the-surprising-way-ebay-used-big-data-analytics-to-save-millions/

[101] cited in Arthur, op. cit., 2012.

[102] cited in Ibid.

[103] Barwick, op. cit., 2013.

[104] Data Science Series, "How eBay Performs Big Data Research to Create New Insights for its Business", *Data Science Stories,* 2012. http://datascienceseries.com/stories/how-ebay-performs-big-data-research-to-create-new-insights-for-its-business

[105] Ibid.

[106] Data Science Series, op. cit., 2012.

Another benefit resulting from the eBay big data analytics programme is the release of eBay's distributed analytics engine, Kylin, to the open-source community.[107] Kylin provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop to support extremely large datasets.[108] Kylin is used in production by various business units at eBay, and in addition to open-sourcing Kylin, eBay has proposed Kylin as an Apache Incubator project because of its big data specific features, such as its design to reduce query latency on Hadoop for 10+ billion rows of data.[109] Kylin was necessitated by the ever-increasing growth of data stores held by eBay and the diversity of the user base. Users in analytics and business users ask for minimal latency but want to continue using tools such as excel and Tableau. As none of the available tools met eBay's exact requirements, eBay built this platform from scratch to meet their emerging business needs and is representative of a situation where big data stores lead to innovative technological developments. eBay big data analytics drove the adoption of infrastructure that is itself a positive experiment for such online companies, especially as the eBay employs a dynamic, ever evolving approach to its relationship with the big data it collects. Although, the theory behind Kylin is not new, it includes methods to store pre-calculated results to serve analysis queries, generate each level's cuboids with all possible combinations of dimensions, and calculate all metrics at different levels.[110] What this means for eBay, and other ecommerce companies is that, when data becomes bigger, the pre-calculation processing becomes impossible even with powerful hardware, unless software like Kylin is implemented to perform these calculations in parallel and merge the final result, thereby significantly reducing the processing time.[111] An example provided by eBay is the storing of several records in hive tables that represent a relational structure. When the data volume grows very large – 10+ or even 100+ billions of rows – a question like "how many units were sold in the technology category in 2010 on the US site?" will produce a query with a large table scan and a long delay to get the answer. Since the values are fixed every time the query is run, it makes sense to calculate and store those values for further usage. This is a powerful development in commercial big data analytics that has resulted from the eBay data analytics programme. Relevantly, at the time of open-sourcing Kylin, eBay already had several business units using it in production. eBay's largest use case is the analysis of 12+ billion source records generating 14+ TB cubes. Its 90 per cent query latency is less than 5 seconds. Now, their use cases target analysts and business users, who can access analytics and get results through the tableau dashboard very easily.[112] This is a good example of competitive companies dealing with open source and entering the open access arena with commercial data. Turning to the future, Liang said that the company was considering the development of machine learning techniques to drive more value from stored data: "You don't need to spend so much time finding different algorithms because once you have a big volume of data, machine learning will offer a higher rate of accuracy."[113] According to Liang, the future will be "live"- meaning real time data loading and analytics. Coupled with forecasting and predicting future events, this will lead to even higher value being delivered by the analytics platforms."[114]

[107] www.kylin.io
[108] eBay Inc., "Announcing Kylin: Extreme OLAP Engine for Big Data", *eBay Tech Blog,* 20 October 2014. http://www.ebaytechblog.com/2014/10/20/announcing-kylin-extreme-olap-engine-for-big-data/#.VEjfF4uUdB2
[109] Ibid.
[110] eBay Inc., op. cit., 2014.
[111] Ibid.
[112] Ibid.
[113] Barwick, op. cit., 2013.
[114] Ibid.

Therefore, eBay's big data analytics programs incorporates a number of elements that make it a meaningful big data story to be examined in line with BYTE which identifies how commercial entities can implement innovative technologies and practices to harness big data to create more efficient business models. In particular, eBay's focus on developing and tweaking big data specific technologies is essential to the success of its initiative. This is particularly so as eBay is currently focussing on growing the open-source community supporting Kylin.

2.1.6 UK Data Service

The UK Data Service is funded by the Economic and Social Research Council (ESRC) of the UK, a major research funding body financed by the UK government. The Data Service holds a wide variety of secondary data resources, including "large-scale government surveys, international macrodata, business microdata, qualitative studies and census data from 1971 to 2011".[115] It's primary function is to support researchers, educators, students, local and national policy-makers, charities, foundations, think tanks, businesses and other data owners and users by providing "high-quality social and economic data".[116] The Service was established in 2012 and it currently has 350,000 catalogue records.[117] As a central point of national data management expertise, the UK Data Service is also heavily involved in the ESRC Big Data Network. Consequently, they are navigating a range of challenges and externalities.

The UK Data Service itself is focused on making data available in order to generate positive externalities in terms of research, economics and policy. Specifically, they are focused on the re-use of existing data sets to inform policy, generate new insights and influence debates. This includes providing support, training and guidance for those depositing and re-using research data. It also involves developing infrastructure and support mechanisms for data storage and exchange. The UK Data Service is currently working with academics, national government agencies and statistics organisations and international organisations such as the International Monetary Fund, European Commission and UNESCO in order "to develop statistical systems for aggregate data and to promote established standards for data exchange to improve data infrastructures."[118] As such, the organisation is involved in mechanisms to coordinate stakeholders within the data ecosystem and to establish standards and infrastructures to enable this exchange. In addition, the UK Data Service is also managing open access mechanisms in relation to these data sets, as described in their 2014 annual report.[119] They have three categories of data. Open data, which does not include personal data, and have few usage restrictions. Safeguarded data which is not personal data, but which could be linked with other data sets to enable the re-identification of individuals. Registration and authentication are required to use this data. Finally, controlled data may include personal data, and those interested in re-using it must be approved by the data access committee and must use the data at the Data Service premises. This open access provision is heavily focused on meeting challenges relating to the protection of personal data; however this is more than likely a result of the types of data that are collected by the Service.

[115] UK Data Service, "About us", 2014. http://ukdataservice.ac.uk/about-us.aspx
[116] Ibid.
[117] UK Data Service, *Annual report: October 2012 – March 2014*, 2014, p. 4. http://ukdataservice.ac.uk/media/455259/ukdataserviceannualreport2012-2014.pdf
[118] UK Data Service, *Annual Report*, op. cit., 2014, p. 19.
[119] UK Data Service, *Annual Report*, op. cit., 2014, p. 5.

These expertise have placed the UK Data Service in a central role within the UK data ecosystem, and the Service will be leveraging this expertise by steering the ESRC Big Data Network. The Network is supported by £64 million in funding from the UK government, and it is aiming to optimise data as a resource. Specifically:

> The enormous volume and complexity of data that is now being collected by government departments, businesses and other organisations represents a significant resource within the UK which can be used to the mutual benefit of academic research, organisations and society as a whole.[120]

As such, the Network is intending to use data as a resource to produce particular positive externalities for British society. These externalities include promoting "the future sustainability of the UK research competitiveness, supporting the UK in maximising its innovation potential and driving economic growth."[121] Thus, most of these externalities are focused on economics and innovation.

However, the strategy document also recognises that privacy and data protection concerns are central to optimising this resource. As such, the initiative mentions the importance of safeguarding individuals' identities. Furthermore, it also mentions that the systems developed for storing, linking and analysing this data must by "safe, secure and efficient" in order to build trust among the different stakeholders implicated in these practices, including members of the public.[122] Thus, like the US initiative discussed above, the ESRC Big Data Network also locates public trust as an essential element of capturing the potential positive externalities of big data.

What makes the ESRC Big Data Network stand out from other initiatives examined in this report is the intention to build a resource that encompasses data from so many different types of organisations. The Network will be constructed in three phases. The first phase, the administrative data research network, is focused on enabling access to administrative data from government departments and agencies, as well as other public sector organisations.[123] Phase two is focused on data "routinely collected" by businesses and local governments. Although the available documents do not specify which types of data this may include, the purpose of such a focus is to "undertake research that makes a difference" by "shaping public policies", "shaping wider society" and "making businesses…and other organisations more effective".[124] As above, both these phases are focused on ensuring that individuals are not identifiable in any data that is produced. Finally, the third phase is focused on social media and third sector data. Currently, only basic information about this phase is currently available, as phase three has been delayed due to the fact that "capacity and capability, data creation, curation and access, risk, ethics and governance" issues surrounding social media data and

[120] Economic and Social Research Council, "Big Data Network", 2014. http://www.esrc.ac.uk/research/major-investments/Big-Data/

[121] ESRC, "ESRC Big Data Network - Phase 2", 2014. http://www.esrc.ac.uk/research/major-investments/Big-Data/BDN-Phase2.aspx

[122] ESRC, "David Willetts MP to announce £14 million funding boost for Data Research Centres", *Press release*, 6 February 2014. http://www.esrc.ac.uk/news-and-events/press-releases/29676/david-willetts-mp-to-announce-14-million-funding-boost-for-data-research-centres.aspx

[123] ESRC, "ESRC Big Data Network - Phase 1", 2014. http://www.esrc.ac.uk/research/major-investments/Big-Data/BDN-phase1.aspx

[124] Economic and Social Research Council, "Big Data Network", 2014. http://www.esrc.ac.uk/research/major-investments/Big-Data/

third sector data need careful consideration.[125] This delay demonstrates that although the UK Data Service has significant expertise in this area, and has taken a leadership role in setting standards related to data management, access and preservation, the variety of data sources being pursued by the initiative are such that the solutions developed for economic and social research data are not sufficient to address these issues. The complexity of the solutions will likely have to develop alongside the complexity of the data itself.

This initiative has demonstrated that the use of data that may contain personal data is a significant challenge in relation to big data, particularly as the number and types of data sets that are available expand. Furthermore, it demonstrates that existing solutions for opening certain types of data may not be sufficient for addressing these challenges in relation to high volumes or large collections of diverse data. However, the initiative also demonstrates the importance of working with other organisaitons to develop shared practice and standards and to ensure that the needs of members of the public, in particular, are being adequately addressed. Establishing trust by protecting privacy and the security of personal data is essential to enable big data stakeholders to capture the potential positive externalities associated with this "resource" including economic efficiencies, job creation, new discoveries and other benefits.

2.1.7 European Space Agency Big Data Initiative

The "Big data from Space" is the European Space Agency initiative to address the barriers that hamper an effective use of big data in earth observation[126]. This makes it a meaningful big data story for BYTE, which seeks to address barriers and issues that arise preventing big data actors from capturing the full potential of the big data they hold. The main rationale is that earth observation data are growing in size and variety at an exceptionally fast rate (see Figure 1), posing challenges and opportunities for their access and application.

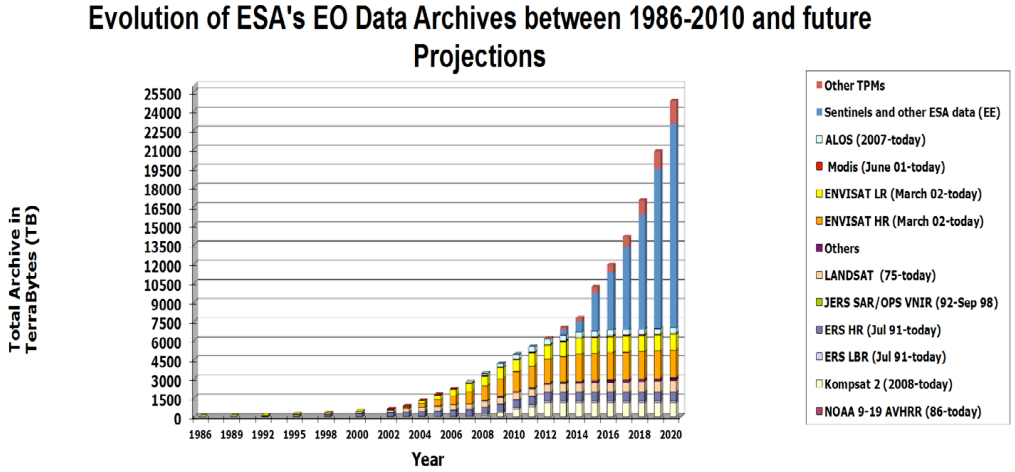

Figure 1 - Evolution of ESA EO data archives[127]

[125] ESRC, "Communication on ESRC Big Data Network Phase 3", 20 December 2013. http://www.esrc.ac.uk/_images/Communication-on-ESRC-Big-Data-Network-Phase-3_tcm8-29496.pdf

[126] Mazzetti, Paolo, Angelo Pio Rossi, Oliver Clements, Stefano Natali, Simone Mantovani, Maria Grazia Veratelli, Joachim Ungar, John Laxton, *Community Reporting*, D120.21 EarthServer Project, 31 August 2013.

[127] Houghton, Nigel. "ESA Reprocessing. A service based approach". Big data from Space Conference, Frascati (Rome), 5-7 June 2013.

ESA has recognised[128] that an unpredictable added value is derivable from the innovative analyses and fusion of big earth observation data, defined in terms of volumes, degree of diversity and complexity, including data streaming from presently available and upcoming satellite capabilities, and ubiquitous ground devices. The increasing diversity and variety of space data, an increasing combined use of diverse space missions data, and an increasing integration of satellite born data with non-space data naturally lead to an unprecedented opportunity to serve new types of user applications, impacting the way such data are collected, referenced, disseminated, processed and delivered. Their combination coupled with today's Internet technologies present new exciting opportunities for generating insights from these high volume and significantly varied data sets beyond the purpose for which the data was originally intended.

Innovative exploitation and the potential economic value and social return are drivers of this big data initiative. Not all the fields of application serve people skilled in understanding earth observation data. Rather, these fields increasingly serve a public focused on issues rather than data, looking for rapid reliable answers to generally complex questions, and bypassing time consuming processes and analyses, which may be necessary in the backend. Generally, people are proficient at finding information when that finding is a reasonable objective, when it is easy to locate the necessary information, and when it is easy to understand the information. Effectively, this means that earth observation data - including those from space - need to be discoverable, machine consumable and handled by interoperable services at the WWW. The emerging Internet of Things will allow combining large heterogeneous environmental information in ways unthinkable just a decade ago, and consumers of such information are readily there. These innovative technologies align with a focus of BYTE on emerging big data technologies and practices.

European space agency big data initiative conference
The first in a series[129] of events organised by ESA as part of this initiative was a conference focused on the issues associated with the organisation and delivery of large volumes of contemporary and historical earth observation data. These data include either space- or ground-based, ubiquitous information-sensing mobile devices, aerial sensory technologies, and wireless sensor networks. The conference, held on 5-7 June 2013 at the ESA's Centre for Earth Observation (ESRIN) located in Frascati, near Rome in Italy, proposed the following objectives:

- To examine current solutions, practices and role of big Earth data services, and identify a common ground.
- To examine issues associated with data organisation and provision, and the associated costs.
- To identify scenarios of data-intensive services, traditional and innovative with respect to new form of processing, enabling additional information derivable from navigation, analytics and correlation of large Earth data sets, and integration across heterogeneous resources.
- To identify challenges, barriers, opportunities for such scenarios, and attempt to define a baseline of activity to make the identified scenarios actionable.
- To critically review current working methods and approaches with respect to the baseline proposal and its application.

[128] ESA Unclassified, Event Report, *Big Data from Space 2013*, Rome (Italy), 5-7 June 2013.
[129] ESA plans to continue convening events in the future, starting from the Living Planet Symposium held in Edinburgh (UK), 9-13 September 2013.

The event was open to decision-makers and technical representatives from all organisations active in using or delivering large complex data sets of earth observations, including: space agencies and satellite operators; agencies/institutions with any R&D/operational requirement for using large earth data volumes; European industrial operators providing services running large Earth data volumes; earth and computer scientists and professionals, as well as students in those areas. The Ground Segment Coordination Body participated as technical advisor. This representation from big data actors presents a number of potential important sources and contacts for BYTE, especially in terms of the participation in the roadmap.

The programme selected by the scientific committee included about fifty talks and thirty posters, ranging from data mining, access and use policies, data interoperability, performance indicators, etc. European Commission's Directorates General Connect, Enterprise and Industry, Research and Innovation, and representatives from the European Environment Agency (EEA), National Oceanic and Atmospheric Administration (NOAA) and Open Geospatial Consortium (OGC) were chairs to the program sessions. The selected oral presentations were organised in four areas: Examples of Application Scenarios, (Big) Best Practices, Examples of Infrastructure Tools, and Some Views in Prospect.

Outcomes
Some 250 scientists, industry representatives, national delegates from Europe, United States, Australia, China and Africa convened at ESA/ESRIN to attend the Big Data From Space event. The event served to stimulate discussion between the different communities involved in the business of providing and manipulating very large-scale data and complex analyses of Earth observations. Selected talks provided a big data vision, critically covering aspects like typical order of data volumes, challenges of data access, including timeliness, policies for their dissemination, data capture, search, sharing, transfer, mining, analysis and fusion, and visualisation issues. Vertical examples touched a broad range of application domains, including: situational awareness, maritime and land cover monitoring, oceanography, climatology, meteorology, and geology. Questions and discussion topics were mainly focused on attempting to identify the state of the art experience and lessons learned; which instruments presently lack to allow effectively capturing and understanding the value of heterogeneous large-scale Earth data sets; which barriers hamper a most effective use of Earth Observation data, space and ground based; concerted solutions, traditional models and businesses compared to new ones.

The main outcomes of the event, grouped per lines of action, can be summarised as follows:

- Challenges
    - The concept of big Earth Observation data is commonly agreed as strictly relating to the capability of rapidly and reliably accessing and consuming them, on one side, and that of running performing analysis and visualisation to discover trends and phenomena otherwise not accessible. Big data applications are by nature borderless, regional and global: big data centres shall cooperate to ensure data are available to such applications.
    - Large amount of heterogeneous datasets require addressing not only Volume but also Variety. E.g. GEOSS is serving 15 million of very heterogeneous resources with 400,000 requests/year.
    - The data deluge will make it increasingly difficult to find and use data of relevance to a given issue. Earth Observation satellite data are much bigger than before, mass processing is more and more required, interdisciplinary

competences are increasingly required. The above requires innovative solutions, more suitable and flexible than the traditional data download at user's end element (also considering that there are still regions with limited bandwidth). They shall naturally lead to setting up downstream services for mass processing or for running sophisticated ad-hoc algorithms.
    - A need to support the ever-increasing data volumes being produced by the next generation of satellite platforms. One speaker reported the need to make a multi-petabyte dataset accessible; another needs to manage 500GB/day, with long-term preservation and curation; others need to support workflow (pre-processing, selection, mosaicking) on TB dataset collections; world data at 12 meter of resolution are already available; lidar data (20 points/m) available in many areas; applications on real-time moisture calculation manages several TB/day of radar backscattering data; EEA manages ~ 30.000 measures per day per pollutant (O3, PM10, PM2.5, NO2, etc.); TomTom manages 6 trillions of measurements since 2008 (6 billions/day now).
- User perception
    - Taking the environment to users, bringing processing to where data are. Physical data aggregation or web processing services approaches; large data holdings associated to commercial hosting, private or public Clouds. Cloud-based platforms are proposed as facilitating the production of qualified EO services.
    - One of the requirements found by Google is the general need for ad-hoc processing moving beyond the portal approach: users want to process data on the cloud not just access them. Download is no more an option; therefore there is the need for user-oriented processing platforms. Acquired and generated datasets should be shared through URLs.
    - With the current solutions available, it is perceived that outreach and training to use data facilities are needed.
    - Finding data correlations often requires cross disciplines expertise. Rise of the data scientist role. Data science will complement traditional domain knowledge.
    - Users may need focussing on issues rather than data, looking for performing visualisation of trends from the most large and reliable basis of information possible (hosted remotely).
- Learning curve
    - User platforms, web tools and crowdsourcing models to structure mass collaboration to environmental monitoring are still at their infancy. Innovative technical solutions often rely on unstable poorly documented technology. Systems scalability and services operationalisation are generally limited by lack of resources.
    - Technical development is envisaged in: data automatic acquisition, discovery and aggregation (cubes); intensive visualisation; security of data access, manipulation, transfer; data and algorithms (standard, machine-readable) description; peak loads enabled web services; spatial processing in Cloud environments; interoperability standards for Cloud based environments; archive policy; semantic mining of information.
- Way forward
    - The wide scope and response given by the many received contributions highlights a high potential to address complex issues via new public-private

partnerships. To this end, further characterisation of partnership models leading to successful sustained businesses is needed.
    - It is common perception that future big data initiatives will trigger a number of activities with favourable economic impact, freeing resources untapped so far.
    - Relying on decades of experience, consolidated processes and a high degree of coordination among its key players, Earth Observation data assets certainly represent a major case to pursue further in the big space data domain.
    - "Geo-spatial is not geo-special", therefore solutions developed for large datasets without any explicit space-time reference might be re-used for big geospatial data. It is what Google itself is doing with the Google Earth Engine platform.
    - Big Data will be an increasing issue since users wish to climb up from L4 products to L0 (raw) measurements. Emergence of data-intensive science. Emergence of crowdsourcing.
    - Need of supporting semantics, also to address data diversity (variety). Need to handle unstructured data, with innovative approaches for data management: no-SQL databases (e.g.BigTable). Need to move from Big Data to Big Information. Innovative approaches for distributed data management (e.g. Hadoop).

A number of conclusions were drawn in relation to the aforementioned outcomes. These conclusions are summarised as follows:

- Many different applications currently need to manage either large datasets (Petabytes not uncommon for EO products) or large amount of small-medium size datasets (as in GEOSS for multidisciplinary applications). Volume and variety appear to be the most cited issues for Big Data Management and analytics. Long-time preservation and data curation are emerging issues.
- There is a strong call for the ability to handle and use big earth observation data, by the most disparate profiles in the public attending the event. More in terms of data diversity and so combined analysis, than volumes. This brings about new opportunities for research, requires new cooperation schemes – European and worldwide – including sound technical, programmatic and industrial coordination, new approaches and new types of service, new skills. Overall, a shift towards utilising more earth observation data is perceived as necessary, e.g. looking wider at what other information areas do.
- The spectrum of potential application use scenarios is wide, touching traditional as well as new application domains, with different scopes and objectives.
- Big Data challenges exist for both products delivery (with consolidated algorithms to calculate essential variables) and research applications (with innovative algorithms and procedures for finding new indicators and information). To support research applications, ad-hoc processing is needed. Research should take into account a paradigm shift towards data-intensive science. Due to the amount of data, download is no more an option. There is the need of taking computation to the data.
- "Taking the environment to the user, bringing processing to the data" is a concept unanimously supported within the public present at the event. It is perceived as complementary to web services based approaches (service oriented architectures remain widely used and promoted). User platforms are conceived primarily to allow users focussing on productive work, e.g. development of new algorithms and services, automating as much as possible time consuming and less productive tasks like data discovery, aggregation, and processing.

- It is expected that the need will grow due to both sensor enhancements (better spatio-temporal and radiometric resolution), number of sensors (including crowdsourcing) and users' requirements (from access to high-level products, i.e. L4, to intermediate products or even L0 raw data).
- Several innovative technologies exist and are candidate to support Big Data management and analytics:
  - Cloud computing at every level (IaaS, PaaS and SaaS).
  - Distributed data management systems like Hadoop.
  - No-SQL database approaches like Google BigTable.
  - Datacube approach.
  - Unstructured data, linked data and Semantic Web solutions.

Summary
The first event of the Big Data from Space initiative provided insights about the current perception of big data and big data analytics in particular for earth observation applications. This initiative provides a number of useful actionable matters that would enhance technologies and practices that utilise earth observation data to effectively harness the results of big data analytics in this area for scientific discover and further technological development. technological innovation in this area will likely develop as a result of the volume and variety of earth observation data available.

2.1.8 European Bioinformatics Institute

The European Bioinformatics Institute (EBI, launched in 1992)[130] is a non-profit, intergovernmental organisation funded by the European Molecular Biology Laboratory (EMBL)[131], which has 21 Member States. They have over 500 members of staff and it is located on the Wellcome Trust Genome Campus in Hinxton, Cambridge in the United Kingdom.

As one of the research units of EMBL, Europe's flagship laboratory for the life sciences, EBI constitutes a leading institution for storage and analysis of large biological datasets. EBI provides freely accessible data from life science experiments, bioinformatics services built on top of these datasets, as well as training to scientists interested in bioinformatics. Even though this initiative is mostly approaching big data as a technology provider, the socio-ethical implications of the life science datasets they work with are of interest for BYTE, especially on the health case study that will be analysed within the project.

EMBL-EBI services around big data are offered following these basic principles:
- *Accessibility:* Both the data and tools they provide are available for free without any restriction. However, in order to access potentially identifiable human genetic information research consent agreements have to be accepted.
- *Compatibility*: All services are built considering established standards in bioinformatics. EMBL-EBI is actively involved in the promotion and development of those standards in order to foster their adoption and accessibility to data.

[130] The European Bioinformatics Institute, "Home", 2014. http://www.ebi.ac.uk/
[131] European Molecular Biology Laboratory, "Europe's flagship laboratory for the life sciences", 2013. http://www.embl.org/

- *Quality:* Datasets are annotated in order to provide context and ease the interpretation of the original data. These annotations are added both automatically and under the supervision of highly qualified biologists.
- *Portability:* Most datasets can be downloaded from the EMBL-EBI website. Furthermore, in some cases even the entire software system can be downloaded and installed locally.
- *Comprehensive data sets:* EMBL-EBI resources are comprehensive and up to date. The Institute works with publishers to ensure that biological data are available in public repositories so that it can be cross-referenced in the relevant publication.

The successfull application of those basic principles involve a series of measures that are aligned with many actions that will be undertaken in BYTE. Thus, several challenges are posed by bioinformatics research nowadays, because of the ever growing amount of data produced from life science experiments. EMBL-EBI bioinformatics services provide efficient means to store, retrieve and analyse these data.

The wide range of services provided by this initiative includes several areas related to bioinformatics and life science:

- DNA & RNA data archival and analysis tools, including genes, genomes and variations from different species.
- Gene, proteins and metabolites expressions databases.
- Protein sequencing datasets and search tools for those databases.
- Structural data at molecular and cellular levels.
- Curated databases of biological systems, to analyse reactions, interactions and pathways within them.
- Chemical biology services, including datasets about chemogenomics and metabolomics.
- Ontogies, taxonomies and controlled vocabularies about life science.
- Literature databases, including scientific publications, patents, and other research objects.
- Other cross-domain tools, software and resources that can be used in life science research.

All these services can be accessed programatically, allowing users to develop data analysis pipelines or integrate public data with their own applications. The complete catalog of web services[132] provides both SOAP and REST interfaces to access them. Some examples of EBML-EBI hosted, publicly open, and free to use life science resources are the following:

- ArrayExpress - archive of gene expression experiments
- BioModels Database - a database of computational models relevant to the life sciences
- Chemical Entities of Biological Interest (ChEBI) - database and ontology of molecular entities
- European Nucleotide Archive (ENA) - resource of nucleotide sequencing information
- Ensemble project - genome databases for vertebrates and other eukaryotic species (joint with Wellcome Trust Sanger Institute)
- Europe PubMed Central - database offering free access to collection of biomedical research literature

[132] Euorpean Bioinformatics Institute, "EBI-EBML Web Services", 2014. http://www.ebi.ac.uk/Tools/webservices/

- Experimental Factor Ontology (EFO) - ontology of experimental variables for biomedical data
- Expression Atlas - database of summary information on which genes are expressed under which conditions
- Gene ontology - ontology of gene functions and processes
- Protein Data Bank in Europe - European resource for the collection, organisation and dissemination of data on biological macromolecular structures
- Proteomics Identifications Database (PRIDE) - repository of mass spectrometry (MS) based proteomics
- UniProt - database of protein sequence and functional information (joint with Swiss Institute of Bioinformatics and Protein Information Resource)

Apart from these services, EBI also offer a collaboration infrastructure accessible by project partners so that they can directly access their datasets hosted at EBI and the institute's powerful computing resources. This cloud computing infrastructure, named The Embassy Cloud, have been successfully piloted within several projects, such as Europe PubMed Central and Tara Oceans, and it is now widely available for other external and collaborative projects.

Another major area within EMBL-EBI activities is training. They offer an extensive user-training programme, aimed at researchers and practitioners in the bioinformatics community. Courses are conducted on site, at interested host institutions around the world, and they are also offered through an online platform.

Finally, been an active stakeholder in big data in Europe for more than 20 years, EBML-EBI has been an important advocate for European coordination and collaboration in bioinformatics and life science research. They play a central role in several European research infrastructures, such as ELIXIR[133], the emerging pan-European infrastructure for biological information. The objective of ELIXIR is to support life science research and its application to medicine, agriculture, bio-industries and the society in general. It is not only focused on providing facilities for life science researchers, but it is also taking care of managing and safeguarding the large amounts of sensitive data produced by integrated publicly funded research projects. In this regard, privacy externalities as defined in BYTE are present. Related to ELIXIR, EBML-EBI also participates in BioMedBridges[134], a project that focuses on interoperability between data and services in the ESFRI biomedical sciences research infrastructures.

Like the other initiatives examined in this report, the EBI's data collection and use policies includes a number of different elements relevant to BYTE. First, the focus on open data is particularly relevant for BYTE, as the availability of data sets is a key facet of ensuring access to large volumes of data. Furthermore, the EBI manage this by outsourcing ownership of this data to the original researchers in order for them to determine which data are sensitive, and which can be made openly available. This practice is somewhat in contrast to the UK Data Service, who manage access to the data after it has been initially classified by the data owner. In addition, the EBI-EBML partnership has also resulted in the creation of an infrastructure to manage the storage and processing of this data on a Europe-wide basis. Thus, the EBI was,

[133] Elixer, "Welcome to Elixer", 2014. http://www.elixir-europe.org/

[134] Biomed Bridges, "Building data bridges from biology to medicine in Europe", no date. http://www.biomedbridges.eu/

and continues to be, particularly engaged in creating a big data set, and enabling the use of that information for a wide variety of scientists in order to generate new insights, discoveries and innovations.

2.1.9 Big data platform vendors

There exists an emerging trend towards collaboration for the development and provision of data platforms. More recently this has included the promotion of open source platforms for big data. Big data platform vendors "provide a specialized product and services offering to address the challenges of big data scenarios."[135] Commercial data platform vendors such as Teradata, Cloudera and the Open Data Platform (ODP) are important because they may be instrumental in the expansion of standardised platforms for big data analytics across a broad range of sectors:

> This comes at a time when big data analytics projects seem to be gaining ground in corporate IT. As an instance, big data analytics and management emerged as a stronger area of focus in the annual TechTarget/Computer Weekly IT Spending Priorities survey for 2015 than in 2014. 30% of respondents globally said they were undertaking big data initiatives in 2015, while the figure for Europe was 26%, and for the UK 21%. The global figure for the previous year was 17%."[136]

Commercial vendors, such as Teradata, Cloudera and the Open Data Platform can either support the growth of industry or perpetuate the existence of closed data sets depending upon accessibility of their models to all stakeholders. Teradata[137] is a commercial enterprise that provides tools and platforms for the implementation of big data initiatives. It specialises in data warehousing and big data analytics. Cloudera is an American-based software company that provides Apache Hadoop-based software[138], support and services, and training to business customers. The Cloudera business model is based on the premise that enterprises strive to be information-driven and make decisions based on data, rather than hunches. The ODP is a shared industry initiative. A group of IT suppliers launched an "open data platform" association to boost big data technology and promote the adoption of Hadoop. These three examples of open data platform vendors represent collaborative (and at times open) data platforms that aim to maximise adoption and productivity of big data utilisation. Whilst they are not without criticism by stakeholders, especially with respect to access and registration structures, the platforms are industry neutral. This means they provide interoperable means for stakeholders to implement data initiatives (especially initiatives dependent on real-time data analytics) across key sectors, including energy, transport, environment, health, culture, crisis informatics and utilities/ smart cities.

---

[135] Landrock, Holm., Oliver Schonschek and Prof. Dr. Andrea Gadatsch, *Big Data Vendor Benchmark 2015: A Comparison of Big Data Solution Providers*, Experton Group, Germany, 2015, p.6. http://www.t-systems.com/solutions/big-data-vendor-download/1298900_1/blobBinary/Big+Data+Vendor_Download-ps.pdf

[136] McKenna, Brian., "Open Data Platform industry association launched to promote Hadoop", *ComputerWeekly,* 26 February 2015. http://www.computerweekly.com/news/2240241295/Open-Data-Platform-industry-association-launched-to-promote-Hadoop

[137] Teradata was founded in 1979 in the United States. It is now present in 42 countries, with more than ten thousand employees and 2.5 thousand customers in 77 countries

[138] Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

The Platforms
The platforms provided by the three big data platform vendors discussed in this report, Teradata. Cloudera, and the ODP, provide platforms for solutions and applications to transform big data sets into valuable knowledge for informed decision-making. The platforms aimed at big data analytics are varied.

First, Teradata provides general purpose platforms as well as specific business applications to translate big data into business intelligence. These include: the Unified Data Architecture, the Aster Discovery Platform, Data Warehouse Software; Data Platform Software, Work-load Specific Platforms; and Teradata Unity that offers a complete portfolio of integrated big data products. Second, Cloudera focuses on providing a united platform that supports information-driven decision-making. This can in turn support big data based innovation. Its partner ecosystem[139] build applications upon core Hadoop technologies. [140] Cloudera says that more than 50% of its engineering output is donated upstream to the various Apache-licensed open source projects (Apache Hive, Apache Avro, Apache HBase, and so on) that combine to form the Hadoop platform. Third, the ODP aims to promote big data technologies based on open source software from the Apache Hadoop ecosystem, as well as optimise testing among and across the ecosystem's vendors.

All three big data platform vendors build upon Hadoop, which has been widely adopted in the global digital market. For example, Facebook utilised Hadoop for building analytic applications involving massive volumes of user data.[141] These platform vendors' approach to big data offers adaptable and advanced technologies to deal with storage, processing and analytics needs of large datasets in multiple domains.

The data
All platform vendors focus on any data that is big data. The data source is only relevant in so far as the platforms are interoperable. Otherwise, the platforms can process big data from data sources across any industry or sector. The BYTE case study areas of the European sectors of energy, transport, culture, health, crisis informatics and environment and utilities/ smart cities all produce data that is compatible with these platforms. In that sense, the platforms are data neutral. This is, amongst other benefits is what makes these platforms stand alone big data initiatives, whilst enabling and supporting the expanding big data economy and its stakeholders in other big data-driven initiatives.

The benefits
Commercial big data platform vendors are driven by their own business needs and interests, as well as being the drivers of collaboration, innovation and standardisation in big data technologies, infrastructures, and practices. The platforms they provide support innovation in the area of big data platform technology and industry as well as supporting companies and organisations that adopt them with opportunities for internal data-driven innovation and decision-making. In the instance of the ODP, a key benefit will be "for members to collaborate across various Hadoop Apache projects as well as other open source-licensed big

[139] Cloudera's partner ecosystem includes some 1,447 companies.
[140] Twentyman, Jessica., "Open Data Platform: the Answer to a Question that No One Asked?", *Computer Weekly,* April 2015. http://www.computerweekly.com/feature/Open-Data-Platform-answer-to-question-no-one-asked
[141] "Cloudera", *Wikipedia,* no date. https://en.wikipedia.org/wiki/Cloudera

data projects with a goal towards meeting enterprise class requirements".[142] Further, the ODP association promotes 8 key features of its platform that can loosely translate across all three big data platform vendors discussed in this report. These key benefits are:

1. Accelerate the delivery of big data solutions by providing a well-defined core platform target;
2. Define, integrate, test and certify a standard ODP Core of compatible versions of select big data open source projects;
3. Provide a stable base against which big data solutions providers can qualify solutions;
4. Produce a set of tools and methods that enable members to create and test differentiated offerings based on ODP Core;
5. Reinforce the role of the Apache Software Foundation (ASF) in the development and governance of upstream projects;
6. Contribute to ASF projects in accordance with ASF processes and Intellectual Property guidelines;
7. Support community development and outreach activities that accelerate the rollout of modern data architectures that leverage Apache Hadoop; and
8. Will help minimise the fragmentation and duplication of effort within the industry.[143]

There are of course a number of benefits for enterprise in terms of generating growth and revenue. With respect to the extent of the contribution Cloudera can make to enterprise, its website provides:

> Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.[144]

Moreover, the big data platforms are also encouraging open source platforms and technologies, which may move the platforms towards a more accessible model for a wider range of organisations, rather than just those who are current subscribers. This has obvious benefits for European, and indeed the global, big data industry. In the instance of the ODP, its business model is based on an open-source model, with the software it freely distributes being based on ASF-developed code. Initially, the ODP members will focus on developing and using offerings of the Apache Hadoop use cases. This will support the ODP in providing access to a "tested reference core of Apache Hadoop, Apache Ambari and related Apache source artifacts."[145] The effect of this is said to simplify upstream and downstream qualification efforts, thereby giving the industry a "test once, use everywhere" core platform.

Furthermore, the benefits extend beyond those that flow from adoption of the big data platforms. Other benefits have been produced by the vendors that encourage and facilitate the

[142] Lee, Michael and Alex Plant., "Technology Leaders Unite Around 'Open Data Platform' to Increase Enterprise Adoption of Apache Hadoop and Big Data", *Pivotal press release,* 17 February 2015. http://pivotal.io/big-data/press-release/technology-leaders-unite-around-open-data-platform-to-increase-enterprise-adoption-of-hadoop-and-big-data

[143] "Open Data Platform", no date. http://opendataplatform.org

[144] "Cloudera's leadership in the Apache Hadoop Community Drives Accelerated Enterprise Adoption and Strong Business Results in Fiscal 2015" *Cloudera,* 17 February 2015. http://www.cloudera.com/content/cloudera/en/about/press-center/press-releases/2015/02/17/cloudera_s-leadership-within-the-apache-hadoop-community.html

[145] Lee, Michael and Alex Plant., op.cit, 2015.

integration of big data platforms across all industry. For example, Teradata is involved in research initiatives on big data, publishing research articles and white papers on big data, providing insights on recent trends in big data and its adoption by businesses from different sectors. Teradata has recently published a white paper[146] that offers recommendations for organisations that want to apply big data vision to their data architecture. Whilst this ultimately promotes use of their platform and enhances their business potential, these activities are still valid contributions to the big data ecosystem.

Therefore, BYTE can benefit from the applied knowledge of these initiatives to enhance the analysis of externalities present in real applications of big data technologies. The big data platform vendors' offerings are indicative of the growing demand for a united platform to enable businesses to extract value from the data they hold.

The barriers

Whilst there are benefits in terms of innovation in the platforms and the business and organisational benefits for organisations that adopt the platforms, commercial big data platform vendors can potentially perpetuate issues associated with inequality of access. Stakeholders that are not in a position, either in terms of funding or human skill and resources, to capture the benefits that flow adopting these big data platforms may be left behind. The vendors are commercial entities and require subscription and membership fees to sustain their businesses. However, these fees may not be viable for all stakeholders.

Furthermore, Teradata, Cloudera and the ODP are in themselves and jointly with other members amongst the biggest and strongest big data analytics and platform companies in the world. These companies already have sizeable market shares. For example, the ODP members include GE, Hortonworks, IBM, Infosys, Pivotal, SAS, and Altiscale, Capgemini, CenturyLink, EMC, Teradata, Splunk, Verizon and VMware. This is relevant when these already dominant big data players are collaborating to standardise the industry. Smaller players, are unlikely to be competitive with larger big data platform vendors, which means they will have little influence in the standards that will be set for them, and potentially the big data platform industry as a whole.

However, this may be counteracted by the fact that the platform vendors build upon the same Hadoop model. Gartner Analyst, Nick Heudecker, observes, "We're talking about technologies that are basically built on the same stack, so porting from one to another, you might have friction at the management console level, yes, but beyond that, there are a lot of similarities between suppliers."[147] Nevertheless, Navin Budhiraja, Head Architecture and Technology at Infosys observes that in the instance of the ODP ecosystem, it "will require them to deploy new web-scale architectures, and the adoption of these modern architectures can be greatly accelerated if they are based on open standards, and easy access to trained talent."[148]

Ultimately, what this may suggest is that a combined effort may further reduce fragmentation of platforms by creating a more open environment, which supports access and in turn,

[146] Woods, Dan, and Scott Gnau, "How to Stop Small Thinking from Preventing Big Data Victories", October 2013. http://www.teradata.co.uk/Resources/White-Papers/How-to-Stop-Small-Thinking-from-Preventing-Big-Data-Victories/?processed=1&LangType=2057&LangSelect=true

[147] Twentyman, Jessica., op.cit., 2015.

[148] Lee, Michael and Alex Plant., op.cit., 2015.

increases the opportunities for innovation. This model would however reduce benefits often associated with competitiveness.

Teradata, Cloudera and the ODP are examples of big data platform vendors that purport to support real-time and actionable insights, self-service exploration, and fluid data schemas to quickly adapt to the dynamic business needs. Nevertheless, the platforms are not entirely devoid of creating barriers to utilisation of big data within the big data industry. However, the issues and barriers raised in this report indicate ways in which different actors in the big data ecosystem may need to work together and leverage one another's expertise in order to identify big data solutions.

Summary
Analytic insights from big data are transforming sectors such as energy, healthcare, transport, culture, environment, crisis informatics and others. Big data platform vendors are enabling customers to take full advantage of the productivity and efficiency gains made possible by these data, subject to stakeholders' abilities to "keep up" with these advancements. Nevertheless, Teradata, Cloudera and the ODP are three examples of big data platform vendors that are driving collaboration, innovation and standardisation in big data platforms and technologies. This in turn, enables and supports big data initiatives, whilst at the same time indicates a potential for issues associated with the emergence of commercially-driven data platforms and standards in the big data industry.

Whilst each offering discussed above has slight differences, similarity lies in their motivation for interoperability and standardised technologies and infrastructures for big data utilisation, particularly in terms of building upon the Hadoop model. This means that commercial big data platform vendors are well placed to support a degree of interoperable big data utilisation for innovation and productivity. However, access to these platforms may be prohibited by membership and subscription fees, as well as being subject to a company or organisation's ability to attract the requisite level of expertise in their employees.

Nevertheless, the early expertise generated by market leaders such as Teradata, Cloudera and the ODP can inform the policy issues and analytical strategies and these are key stakeholders with whom the BYTE project should engage to leverage these expertise, especially in attempts to pursue a more open-source, rather than proprietary approach to big data. Open-source platforms may create an ecosystem that preserves innovation in open-source and open data, as well as providing vendor support and interoperability for the implementation of big data initiatives. Moreover, the platforms and activities of the three big data platform vendors discussed in this report are of particular interest to BYTE as a source of relevant information pertaining to a variety of big data initiatives supported by big data platforms and technologies as implemented across a variety of sectors by a number of different stakeholders. In essence, big data platform vendors are commercial initiatives in big data that have harnessed knowledge of big data analytics and big data technologies and processes to enable the implementation of future big data initiative.

2.1.10 *deCODE-genetics* analysis, Iceland

The Icelandic initiative deCODE-genetics ("deCODE") is a scientific initiative that specialises in analysing and understanding the human genome through analysis of relevant big data. deCODE genetics is a commercial scientific enterprise foremost, although it has received public support for its advances in genome-related discoveries. deCODE is a valuable

big data story that displays a number of the benefits that flow from big data analytics, primarily, scientific research and developments that aid health and science in society. This is achieved by the following process: "Finding genetic risk factors for disease requires the ability to correlate two large sets of data: on variations in the sequence of the genome on the one hand, and on variations in phenotype or condition, such as a disease or some physical trait, on the other."[149] Thus, deCODE provides an interesting example of a big data initiative relevant to BYTE because it undertakes specific data processes that are dictated by the nature of data being processed.

The data

The data used in this initiative is in the form of a large dataset of genomics data acquired through population studies. deCODE currently analyses genetic and medical data from 500,000 people from around the world, including 140,000 people from Iceland (or more than half the adult population).[150] deCODE is seeking DNA from up to one-third of the total Icelandic population. This DNA is considered essential to future big data analytics carried out by deCODE. This is because the Icelandic population, comprising of approximately 320,00, is so homogenous that the DNA contains insights into what causes specific diseases.[151] Thus, the data held and analysed by deCODE is of particular significance in terms of disease identification and possible prevention. Analysis of such data can produce untold results, and results that are otherwise simply not available outside of a homogenous population similar to Iceland's. However, while the volume of data analysed by deCODE is certainly large, the velocity and variety aspects of the data do not appear to be in evidence. As such, this initiative, like many others appears to challenge the traditional and often-citied Gartner definition of big data.

Results

The analysis of that data has led to the discovery of several genetic risk factors for common diseases. The analytics is carried out in a privacy-friendly manner to minimise negative externalities that can flow from the use of personally identifying data. deCODE utilises statistical algorithms, software programs and sample handling and privacy protection systems that they have developed in-house. This is particularly relevant to BYTE, which addresses big data specific technologies and practices and can learn from the tools and technologies developed by deCODE. The process undertaken and the potential results and their impacts are described:

> The common diseases – such as heart attack, asthma, stroke and cancer – arise from the interplay of multiple genetic and environmental and lifestyle factors. Moreover, our ability to genotype, or read specific 'letters' in the genome, is constantly and rapidly increasing. It is now possible, though still costly, to sequence entire human genomes relatively swiftly. This will enable an ever more detailed understanding of the rare as well as common variations that impact disease, but will require the ability to analyze all 3 billion letters in the genome for every individual studied, compared to the hundreds of thousands of single letter variations, or SNPs, measured by current SNP chips.[152]

[149] deCODE - Genetics, "Unrivaled Capabilities", *Science,* no date. http://www.decode.com/research/
[150] Ibid.
[151] Sigmundsdottir, Alda, "Privacy on Ice: This Company Wants to Collect DNA from one-third of Iceland's population", *Slate,* 21 May 2014.
http://www.slate.com/articles/technology/future_tense/2014/05/decode_genetics_wants_to_collect_dna_from_one_third_of_icelanders.html
[152] deCODE - Genetics, "Unrivaled Capabilities", *Science,* no date. http://www.decode.com/research/

Their expertise with generation and analysis of genetic and medical data on a large scale also means results in effective DNA-based diagnostics for common diseases, as well as personal genome analysis services.

Concerns
deCODE raises the ethical and legal concern of privacy of personal data included in the large data stores analysed by the initiative. BYTE has already identified a number of related issues in relation to big data analytics in the health and science sectors as they deal largely with sensitive personal data about their subjects. In that regard, deCODE's data collection techniques generated a huge controversy when the Icelandic government permitted deCODE's creation of a health database with genomics, genealogical, and health records of citizens. The ethical and legal issues, particularly privacy, were again implicated when deCODE contacted approximately 100, 000 people to provide samples of their DNA to enter into deCODE's databases for prospective studies in relation to DNA factors that cause specific diseases.[153] The social and ethical issues raised by deCODE's collection and subsequent practices with the big data they collect and use underlines the importance of developing ethical and compliant big data practices so that negative externalities are minimised, and society and organisations remain in a position to extract meaningful knowledge and value from big data.

Therefore, deCODE provides an interesting and relevant big data story for BYTE. The particularity of the data at the centre of the initiative is such that it produces results that may not be gleaned from other large data sets in the field of genomics. Furthermore, deCODE has designed and utilised its own technologies for big data analytics, not only to extract the full potential of the data, but to also minimise negative externalities commonly associated with scientific and health research, such as risks to privacy. This level of innovation in practice and technology is an important aspect of BYTE in terms of assessing the potential of big data initiatives both now, and into the future to support the emerging big data industry.

[153] Kirby, Emma Jane, "Iceland's DNA: The World's Most Precious Genes?", *BBC news online,* 18 June 2014. http://www.bbc.com/news/magazine-27903831

3 SUMMARY

A number of big data initiatives are emerging around the globe and provide useful examples for BYTE as they reflect a number of the positive and negative externalities addressed by BYTE. The initiatives discussed in this report provide practical examples of where big data is being harnessed to produce benefits across society, whilst simultaneously attempting to minimise any negative externalities produced by such initiatives. We have broadened our discussion to initiatives from third countries such as Australia and the US as a useful way to compare big data initiatives in Europe and beyond its borders. This also provides a more global perspective of the initiatives that are being undertaken that relate specifically to big data, big data practices and technologies.

The US Big Data Research and Development initiative, the Australian Government Public Service Big Data Strategy, Global Pulse and the ESRC Big Data Network (managed by the UK Data Service) are sound examples of big data strategies with policy support and clear action plans for government's to capture the benefits of big data for the socio-economic and potentially policy gains. Specifically, these initiatives seek to improve the lives of people on the ground by generating positive externalities in the form of job creation, efficiency, information gathering and evidence-based policy-making. These initiatives may also alter the way in which they are able to partner with stakeholders around the world, as well as enable them to contribute information to the global society. These initiatives provide important insight into how supra-national, European and third country governments are dealing with the deluge of data.

Big data initiatives in the scientific arena including the CERN initiative, deCODE, the UK Data Service, the European Space Agency and the European Bioinformatics Institute initiative represent the way scientific data is now analysed for societal improvements. While each of these initiatives have built their own data collection, storage and analysis infrastructure, these initiatives typically raise ethical and legal concerns, particularly in relation to privacy, because of the sensitive nature of the personal data being analysed for results. Because of this, initiatives are producing new and innovative technological and administrative means of enhancing privacy throughout the analytics process. These include the use of virtual machines, distributed computing networks and firewalls to manage data, including ensuring its quality and veracity. Furthermore, deCODE, the UK Data Service and the EBI all use administrative procedures related to data access and ownership that protect privacy. These innovations are in themselves a positive externality of big data initiatives as identified by BYTE.

eBay Inc. big data analytics program and Teradata are private-sector examples that recognise the growing commercial value of big data analytics. eBay provides an example of a big data driven business model to provide consumer benefits and increase profits, whilst Teradata assists in replicating such business models for organisations ad companies to enable them to derive the maximum value from the data they hold. Each of these examples demonstrate that there are significant skill sets already in evidence in the big data ecosystem, and that these need to be mined and leveraged to enable newcomers to the big data arena to build upon this expertise.

In addition to these sector-specific observations, this analysis also results in a number of important conclusions for big data in relation to the objectives of BYTE. First, the analysis challenges the traditional and long-accepted basic definition based on the 3Vs originally

proposed by Gartner. All of these initiatives are dealing with large volumes of data. However, most are only dealing with either high velocity or data of a large variety. Very few are dealing with all three. Furthermore, this investigation also finds that the conceptualisation of high velocity data varies significantly based on the type of data being utilised. For example, some data involves high velocity collection or high velocity growth, others involve high velocity transfer and still others involve high velocity analysis. Furthermore, velocity is relative and depends on the volume and type of data being processed. The CERN Worldwide LHC Computing Grid offers a particularly interesting example, in that the transfer of data from the core is characterised by enormous transfer velocity requirements, yet the analysis of these exceptional volumes of data generally require days to complete. BYTE D1.1 on *Understanding and mapping big data* will examine this in more detail; however, this analysis suggests that "big data" may be context dependent.

Second, all the aforementioned big data policy initiatives involve cross-sector and cross-agency collaborations, business partnerships and similar emerging relationships. In the Australian initiative specifically foregrounds collaborations and partnerships between governments and businesses in order to achieve the expected benefits of big data. In addition, the US example includes a specific mandate that implicates six different government departments and agencies, while the UK's ESRC Big Data Network includes data sets from government, businesses and third sector organisations that will in turn benefit each of these different stakeholder types. This demonstrates that form a policy perspective, the benefits of big data would best be achieved through different stakeholders working collaboratively. This is also one of the key objectives of BYTE and the Big data community that BYTE will form, and the policy push to create similar networks demonstrates that such networking and community building is understood to represent good practice in responsible innovation.

Finally, and most importantly, the aforementioned big data stories detail initiatives that have been implemented to utilise big data to produce positive results, impacts and externalities and provide initial information about how some big data initiatives are addressing potential negative externalities. Many of the initiatives examined here, including deCODE, EBI and the UK Data Service are examining innovative ways to protect privacy, prevent discrimination and ensure the protection of personal data. In addition, CERN, ESA, Teradata and eBay all demonstrate that significant economic gains can be realised through the leveraging of big data either to create profit, create jobs, develop new skills or support innovation. Finally, Global Pulse and the UK Data Service demonstrate that big data can be used for social gains, including evidence-based policy-making and the protection of citizen safety, health and well-being. BYTE will consider each of these externalities and strategies as the project develops, particularly in the examination of the sector-specific case studies in D3.1.

4 CONCLUSION

This analysis has considered ten different big data initiatives in Europe, the US, Australia and Iceland, including policy initiatives, commercial initiatives and scientific initiatives. The results of this examination suggest three key avenues of further exploration for the BYTE project as it develops. First, BYTE needs to continue to consider the definition of "big data", and the extent to which it might be contextually specific, throughout the project. It suggests that the project may need to adopt a fluid definition of big data that is flexible and adaptable to different sectors. Second, it underlines the importance of stakeholder collaboration to tackle new scientific developments in a way that captures the benefits of the endeavour whilst protecting citizens from potential negative externalities. Finally, the analysis has also resulted in some specific ideas about the potential externalities that may result from the large-scale deployment of big data, as well as some initial evidence about how some organisations have sought to solve these challenges. Each of these aspects will be brought forward into the further work of BYTE as the project develops.